THE BOOK WAS DRENCHED

SCIENTIFIC INFERENCE

Cambridge University Press
Fetter Lane, London

New York
Bombay, Calcutta, Madras
Toronto
Macmillan

Tokyo
Maruzen Company, Ltd

SCIENTIFIC INFERENCE

by

HAROLD JEFFREYS
M.A., D.Sc., F.R.S.

CAMBRIDGE
AT THE UNIVERSITY PRESS
1931

PRINTED IN GREAT BRITAIN

CONTENTS

PREFACE

THE present work had its beginnings in a series of papers published jointly some years ago by Dr Dorothy Wrinch and myself. Both before and since that time several books purporting to give analyses of the principles of scientific inquiry have appeared, but it seems to me that none of them gives adequate attention to the chief guiding principle of both scientific and everyday knowledge: that it is possible to learn from experience and to make inferences from it beyond the data directly known by sensation. Discussions from the philosophical and logical point of view have tended to the conclusion that this principle cannot be justified by logic alone, which is true, and have left it at that. In discussions by physicists, on the other hand, it hardly seems to be noticed that such a principle exists. In the present work the principle is frankly adopted as a primitive postulate and its consequences are developed. It is found to lead to an explanation and a justification of the high probabilities attached in practice to simple quantitative laws, and thereby to a recasting of the processes involved in description. As illustrations of the actual relations of scientific laws to experience it is shown how the sciences of mensuration and dynamics may be developed. I have been stimulated to an interest in the subject myself on account of the fact that in my work in the subjects of cosmogony and geophysics it has habitually been necessary to apply physical laws far beyond their original range of verification in both time and distance, and the problems involved in such extrapolation have therefore always been prominent.

My thanks are due to the staff of the Cambridge University Press for their care and courtesy; also to Dr Wrinch and Mr M. H. A. Newman, who have read the whole in proof and suggested many improvements.

HAROLD JEFFREYS

ST JOHN'S COLLEGE
CAMBRIDGE

January 1931

CHAPTER I

LOGIC AND SCIENTIFIC INFERENCE

"Contrariwise", continued Tweedledee, "if it was so, it might be; and if it were so, it would be: but as it isn't, it ain't. That's logic."

LEWIS CARROLL,
Through the Looking Glass

1·1. The fundamental problem of this work is the question of the nature of scientific inference. The data available to the scientific worker, as well as to the man in the street, are composed of two classes. The first class consists of the crude data provided by the senses. These will be called *sensations*. The second class consists of general principles, which determine how the information provided by the senses is to be treated. It is actually treated in two different ways, which may be called *description* and *inference*. Description, in the strict sense, would involve only the cataloguing and classification of sensations already experienced. Inference is the use of sensations already experienced to derive information about sensations not yet experienced, to construct physical objects, and to describe the past and future of these physical objects. For pure description only an application of the principles of classification and the properties of classes is required; these are purely logical ideas.

Inference requires much more. However fully one's past experience has been described and indexed, nothing not included in it can be inferred without some principles not purely logical in character. As a matter of logic this is a commonplace. Actually one proceeds, in the simplest type of inference, on the supposition that what has been found to be true in previous instances will be repeated in new instances. The distinction between deductive logic and scientific

inference may be illustrated by means of one of the classical instances of the former.

All men are mortal.
Socrates is a man.
Therefore Socrates is mortal.

This type of argument, the syllogism, is one of those chiefly used in pure logic; indeed it was believed for ages that there was no other. The first, or general, statement about all men is called the *major premiss*, the particular statement that Socrates is a man is called the *minor premiss*, and from the two together we draw the *conclusion* that Socrates is mortal. But as a scientific argument it is unsatisfactory. We question immediately whether the major premiss is true. It is not known by experience. We cannot state as a result of experience that a man is mortal until he is dead. At any instant men are living, and they all constitute unverified instances of the premiss; it is simply unknown by experience whether the general statement is true or not. Gulliver, arriving in Luggnagg, might have said equally well:

All men are mortal.
A Struldbrug is a man.
Therefore a Struldbrug is mortal.

But Gulliver knew better and did not argue with his informers.

There are several ways of treating the classical syllogism so as to make it somewhat more acceptable to scientific thought. One is to say that the general proposition is not asserted from experience at all, but is known to be true in all possible cases from previous knowledge. In such a case the syllogism becomes valid. But we avoid the difficulty only by admitting that there may be knowledge applicable to the study of experience and not itself derived from experience. This type of knowledge we call *a priori*. We do not say that it is the solution of the present difficulty, but *a priori* knowledge exists, and we shall have occasion later to consider instances of it at length.

The word *mortal* itself introduces difficulties of a type that will concern us later. Suppose that we accepted the syllogism and that Socrates had nevertheless survived till the present day. We should still not be compelled to reject the conclusion of the syllogism. If a doubter pointed out that Socrates had reached the advanced age of 2000 odd, that would not in the least prevent us from continuing to assert his mortality. Our reply would be that he might die to-morrow as far as the doubter knew; and that would close the matter unless the doubter thought of a new line of attack. But suppose he went on: "You are saying that Socrates will not live for ever. May I point out that even if he lives to be a million years old it could still be said that he would die some day? Your statement has the quality that no evidence could possibly be produced that would contradict it. Even if it is true it still gives no reason to suppose that a man cannot live till he is a million years old. In fact it is vague and unverifiable, and therefore uninteresting". The doubter has at this stage abandoned the attempt to show that the deduction has been falsified by experience; he says instead that it is futile because it is not capable of being compared with experience. This is the scientific attitude.

Both these criticisms of the classical syllogism have analogues in relation to certain modern theories of scientific knowledge, as we shall see later.

1·2. An essential object of scientific inference is to increase knowledge. The syllogism has a place in it just so far as it assists this object. Some syllogisms do; others do not. Consider the following example:

All English policemen are over five feet nine inches in height.
Brown is an English policeman.
Therefore Brown is over five feet nine inches in height.

The syllogism as it stands is perfect. The conclusion is free

from the difficulties of that of the classical syllogism; it is perfectly possible to measure Brown's height. But how do we know the general proposition? If it is known by experience, we have already measured the heights of all English policemen, and therefore we have measured P.C. Brown in particular, and we know directly that his height is over five feet nine inches. The major premiss, in fact, contains the conclusion, and the syllogism tells us nothing that we did not know already. But suppose that we have not made extensive measurements of the heights of policemen, but that we know of the official regulation that no man is appointed to be a policeman unless his height is at least five feet nine. The general proposition is now part of our knowledge without having been verified in all its instances; it is previous knowledge. The inference concerning P.C. Brown is now new knowledge; the syllogism tells us something to expect about his appearance when we meet him that we should not have known without it. Thus the same syllogism may or may not provide new knowledge, according to the means of knowing its premisses.

The syllogism about Socrates raises the same question in a more complicated form. Its author may have had previous intuitive or divinely revealed knowledge, independent of experience, that all men were mortal. If so, he could construct his syllogism and derive new knowledge about the particular man Socrates. But this is not the practical case; belief in human mortality is based on experience. A contemporary of Socrates might proceed in the following way. He would summarize what he knew of the duration of human life. No case was known of a man's having lived for 200 years, and few for 100. This suggests a general rule: all men die before reaching 200 years of age, most before reaching 100. He might look for exceptions among living persons, of whom few were over 100 and none over 200. The general rule was verified with regard to all dead persons, and not contradicted by living ones. It is then stated as a result of experience. The inference concerning

the life of a living person could then be drawn from the rule. It could be said that "Socrates will not live to be 200; he will probably not live to be 100". The fundamental difference between the two methods of approach is that in the former, where the major premiss is known *a priori*, we always proceed from the general to the particular; in the latter we get the major premiss itself by asserting as a general proposition what was previously known only in particular instances. The former method is *deduction*, the latter *induction*. In both cases we proceed from the premisses to the conclusion by means of an apparent syllogism; but there is a significant difference, due to the difference in the nature of the available knowledge about the major premiss. Suppose that two people, while Socrates was alive, both drew the inference that he would not live to be 200, one basing his beliefs on human life in general on intuitive knowledge, the other on previous experience; and suppose that Socrates nevertheless lived to be 2000. Suppose further that our doubter paid a visit to Elysium and interviewed their shades. The former would have to admit that his intuitive knowledge, which he had held with certainty, was wrong, or to say that Socrates was not a man but an immortal god, or perhaps to resort to abuse. The latter would explain that the major premiss in his inference was not known with certainty, but that it was extremely probable on the evidence before him. The inference had been correct for some thousands of millions of people that lived when or after it was drawn, and in the circumstances it was not so bad that there had been one exception to the general rule. If he chose to be aggressive he might ask whether Socrates had been medically examined recently with a view to finding out the causes of his anomalous behaviour; for one of the chief functions of exceptions is to *improve* the rule.

The inference with regard to Socrates has actually been verified, but the situation has arisen with respect to many other scientific laws. At present we are faced with the inaccuracy of Euclid's parallel axiom, which for millennia was

considered intuitively obvious; with the inaccuracy of Newton's law of gravitation, which had been well established by experience and had been believed for centuries to be exact; with the failure in stars of the law of the indestructibility of matter; and with the discordance of the classical undulatory theory of light with the group of facts known as quantum phenomena. For twenty years physical science has been modifying and reconstructing its most fundamental laws as a result of new knowledge. The reconstruction has followed, and will continue to follow, the old *method*, but the results will be different because new facts have to be fitted in. Will modern physics suffer in turn the fate of the old? Perhaps; nobody knows. But in the circumstances we must raise a group of questions more fundamental and general than any physical law. Have recent developments shown that scientific method itself is open to suspicion, and if so, is there a better one? Just how much do we mean when we assert the truth of a scientific generalization? When we have made such a generalization, what reason have we for supposing that further instances of it will be true?

1·3. The answers to these questions may be stated at once. There is no more ground now than thirty years ago for doubting the general validity of scientific method, and there is no adequate substitute for it. When we make a scientific generalization we do not assert the generalization or its consequences with certainty; we assert that they have a high degree of probability on the knowledge available to us at the time, but that this probability may be modified by additional knowledge. Our answer is that returned to the doubter by the second shade. The more facts are shown to be co-ordinated by a law, the higher the probability of that law and of further inferences from it. But we can never be entirely sure that additional knowledge will not some day show that the law is in need of modification. The law is provisional, not final; but scientific method provides its own means of assimilating new know-

ledge and improving its results. The notion of *probability*, which plays no part in logic, is fundamental in scientific inference. But the mere notion does not take us far. We must consider what general rules it satisfies, what probabilities are attached to propositions in particular cases, and how the theory of probability can be developed so as to derive estimates of the probabilities of propositions inferred from others and not directly known by experience.

At the same time a remarkable thing happens. It is found that general propositions with high probabilities must have the property of mathematical or logical simplicity. This leads to a reaction upon the descriptive part of science itself. The number of possible methods of classifying sensations is colossal, perhaps infinite. But the importance of simple laws in inference leads us to concentrate on those properties of sensations that actually satisfy simple laws as far as they have been tested. Thus the classifications of sensations actually adopted in practical description are determined by considerations derived from the theory of inference; and probability, from being a despised and generally avoided subject, becomes the most fundamental and general guiding principle of the whole of science.

CHAPTER II

PROBABILITY

Oh, it ain't gonna rain no mo', no mo',
It ain't gonna rain no mo'!
How in the hell can the old folks tell?
Tain't gonna rain no mo'!

MESSRS LAYTON and JOHNSTONE

2·1. What is probability?

Suppose that a man wishes to catch a train announced to start at 1.0 p.m. When he is a quarter of a mile from the station he looks back and sees that a church clock some distance away indicates 12.55. Will he catch the train?

From previous experience he knows that a quarter of a mile in five minutes means comfortable walking without wasting time. The distance, with slight exertion, can be done in four minutes. Hence he may reasonably expect to catch the train, especially if he hurries slightly. But he has to get a ticket before he will be admitted to the platform. If he finds nobody waiting at the booking office this is a matter of ten seconds; but if there is a queue of ten people it will take two minutes, and he has no means of knowing which will occur in this case. Again, though the church clock is usually reliable, it has been known on a few occasions to be as much as three minutes slow. If that is so on this occasion, and the train is punctual, his chance of catching the train disappears. On the other hand, if the train is a few minutes late, as sometimes happens, he will catch it even if there is a queue and the clock is slow. Further, there is always the possibility of something quite unforeseen, such as an accident on the line. In that event the 11.14 train may arrive at 1.30 and his problem will be solved.

Now we notice that in this situation the man has some definite information, which is relevant to the proposition "he

will catch the train". But numerous other possibilities, none of which he can foresee, are also intensely relevant. Therefore his available knowledge, though relevant to the proposition at issue, is not such as to make it possible to assert definitely that this proposition is true or false. Further, extra data will have a definite effect on his attitude to the proposition. If he meets an astronomer whose watch has just been compared with a wireless time signal, and who assures him that the church clock is accurate, he feels more confident. On the other hand, if a crowded omnibus passes him he expects his worst fears about the queue to be verified. Thus the attitude to the proposition under discussion does not amount to a definite assertion of its truth or falsehood; it is an impression capable of being modified at any time by the acquisition of new knowledge.

Probability expresses a relation between a proposition and a set of data. When the data imply that the proposition is true, the probability is said to amount to certainty; when they imply that it is false, the probability becomes impossibility. All intermediate degrees of probability can arise.

The relation of the laws of science to the data of observation is one of probability. The more facts are in agreement with the inferences from a law, the higher the probability of the law becomes; but a single fact not in agreement may reduce a law, previously practically certain, to the status of an impossible one. A specimen of a practically certain law is Ohm's law for solid conductors. Newton's inverse square law of gravitation first became probable when it was shown to give the correct ratio of gravity at the earth's surface to the acceleration of the moon in its orbit. Its probability increased as it was shown to fit the motions of the planets, satellites, and comets, and those of double stars, with an astonishing degree of accuracy. Leverrier's discovery of the excess motion of the perihelion of Mercury scarcely changed this situation, for the phenomenon was qualitatively explicable by the attraction of the visible matter within Mercury's orbit. Newton's

law was first shown to be wrong, as a universal proposition, when it was found that such matter could not actually be present in sufficient quantity to account for the anomalous motion of Mercury.

The fundamental notion of probability is intelligible *a priori* to everybody, and is regularly used in everyday life. Whenever a man says "I think so" or "I think not" or "I am nearly sure of that" he is speaking in terms of this concept; but an addition has crept in. If three persons are presented with the same set of facts, one may assert that he is nearly certain of a result, another that he believes it probable, while the third will express no opinion at all. This might suggest that probability is a matter of differences between individuals. But an analogous situation arises with regard to purely logical inference. One person, reading the proof of Euclid's fifth proposition, is completely convinced; another is entirely unable to grasp it; while there is at any rate one case on record when a student said that the author had rendered the result highly probable. Nobody says on this account that logical demonstration is a matter for personal opinion. We say that the proposition is either proved or not proved, and that such differences of opinion are the result of not understanding the proof, either through inherent incapacity or through not having taken the necessary trouble. The logical demonstration is right or wrong as a matter of the logic itself, and is not a matter for personal judgment. We say the same about probability. On a given set of data p we say that a proposition q has in relation to these data one and only one probability. If any person assigns a different probability, he is simply wrong, and for the same reasons as we assign in the case of logical judgments. Personal differences in assigning probabilities in everyday life are not due to any ambiguity in the notion of probability itself, but to mental differences between individuals, to differences in the data available to them, and to differences in the amount of care taken to evaluate the probability.

2·2. The mathematical discussion of probability depends on the principle that probabilities can be expressed by means of numbers. This depends in turn on two deeper postulates:

1. *If we have two sets of data p and p', and two propositions q and q', and we consider the probabilities of q given p, and of q' given p', then whatever p, p', q, q' may be, the probability of q given p is either greater than, equal to, or less than that of q' given p'.*

2. *All propositions impossible on the data have the same probability, which is not greater than any other probability; and all propositions certain on the data have the same probability, which is not less than any other probability.*

The relations *greater than* and *less than* are transitive; that is, if one probability is greater than a second, and the second greater than a third, then the first probability is greater than the third. If one probability is greater than a second, the second is said to be less than the first; and if neither of two probabilities is greater than the other we say that they are equal. This postulate ensures the existence of a definite order among probabilities, such that each probability follows all smaller ones and precedes all greater ones.

Such an order once established, we can construct a correspondence between probabilities and real numbers, so that to every probability corresponds one and only one number, and so that of every pair of probabilities the less corresponds to the smaller number. When this is done the system of numbers can be used as a scale of reference for probabilities. But the choice is not yet unique. Obviously if $x_1, x_2, \ldots x_n$ are a set of positive numbers in increasing order of magnitude, $x_1^2, x_2^2, \ldots x_n^2$ are another set, $e^{x_1}, e^{x_2}, \ldots e^{x_n}$ a third, $\frac{x_1}{1+x_1}, \frac{x_2}{1+x_2}, \ldots \frac{x_n}{1+x_n}$ a fourth, and any number of such sets can be found, such that if probabilities correspond term by term with the numbers of one set in order of magnitude they will correspond equally well with those of any other set. We need a further rule before we can decide what number to attach to any given probability. Such a rule is a

mere method of working, or convention; it expresses no new assumption. We decide that

3. *If several propositions are mutually contradictory on the data, the number attached to the probability that some one of them is true shall be the sum of those attached to the probabilities that each separately is true.*

If we do this it follows at once that 0 is the number to be attached to a proposition impossible on the data. For consider any three mutually exclusive propositions p, q, r, and suppose we have the further datum that p is true. The number attached to a proposition impossible on the data being a, it follows that the numbers attached to q and r separately on the data are both a. Hence, by our rule, since q and r are mutually exclusive, the number attached to the proposition that one of them is true is $2a$. But the proposition "q or r is true" is itself impossible on the data and therefore has the number a attached to it. Hence $2a = a$, and therefore $a = 0$.

Again, let us consider any set of m equally probable and mutually contradictory propositions, and call the number attached to any one of them, on the same data, x. If we select any l of them, the number attached to the proposition that one of these l is true is lx, by our rule.

Now take $l = m$, and suppose that on our data there is just one true proposition among the m, but that we have no means of knowing which it is. The number attached to the proposition that one of the m propositions is true is mx. But on our data this proposition is certain, and therefore mx is the number corresponding to certainty, which is a definite constant by Prop. 2. We therefore choose 1 as the constant to be attached to certainty. This is another convention. Thus $mx = 1$, and we derive the rule:

4. *If m propositions are equally probable on the data and mutually contradictory, and one of them is known to be true, each has the number 1/m associated with it. Further, the proposition that one out of any l of them is true has the number l/m associated with it.*

The conditions for the application of this method are practically realizable. Suppose that m balls, one of them with a characteristic mark on it, but indistinguishable by touch, were placed in a bag and shaken. l balls are then withdrawn. Then the proposition that any particular ball is the marked one is inconsistent with the proposition that any other is marked, and all such propositions are equally probable. We have therefore a set of equally probable and mutually exclusive propositions, m in number. Our rule therefore has a practical application. Also m may be any integer, and l may be any integer less than m or equal to it. Hence

5. *Any rational proper fraction, including* 0 *and* 1, *can be a probability number.*

We shall call the class of probabilities expressible by rational fractions R-probabilities.

It follows from this that any probability can be made to correspond to a real number, rational or irrational. For any given probability P either corresponds to a rational fraction or does not. In the former case the proposition is granted. In the latter case every R-probability is either greater or less than P. Hence P divides the R-probabilities into two classes R_1 and R_2, such that the probabilities in R_1 are all less than P and those in R_2 are all greater than P. Also, since the relation "greater than" among probabilities is transitive, every fraction corresponding to an R_2 probability is greater than every fraction corresponding to an R_1 probability. Hence P determines a cut in the series of rational fractions. But this is precisely the method of defining a real irrational number; when it is specified which rational fractions are on one side of the cut and which on the other side, there is one and only one real number that can occupy the cut. We then associate the probability P with this number. In this way we arrive at the result:

6. *Every probability can be associated with a real number, rational or irrational.*

We still have to prove that the results given by our rules

are consistent; that is, if a probability P is greater than another probability Q, that the number associated with P by our rules is greater than that associated with Q. Suppose first that P and Q are both R-probabilities. Then we can find four integers l, m, r, s so that the number associated with P is l/m and that associated with Q is r/s. Now consider a class of ms mutually exclusive propositions containing one true one. We may divide them up into m sets of s each; one and only one of these sets contains the true proposition. The probability-number that one of l of these sets contains the true proposition is l/m. But this is also the probability-number that one of ls propositions selected from the original ms propositions shall be the true one, which by our rule is ls/ms and equal to l/m, as it should be. Thus l/m is the number associated with the proposition that one out of the ls alternatives is true; similarly r/s is associated with the proposition that one out of rm alternatives is true. If then P is greater than Q, the number of alternatives needed to give probability P must exceed that needed to give probability Q; therefore ls is greater than rm. But this is equivalent to saying that l/m is greater than r/s; and therefore the greater probability is associated with the greater number.

Consistency is therefore proved for R-probabilities. For others the result is easily generalized. For if two non-rational probabilities are associated with real numbers a and b, of which a is the greater, we can find a rational fraction l/m lying between them. Then the probability associated with a is greater than that associated with l/m, and that associated with l/m is greater than that associated with b. Hence, in virtue of the transitive property of the relation *more probable than*, the probability associated with a is greater than that associated with b. In other words, the greater number corresponds to the greater probability.

We have seen how definite numbers can be associated with probabilities, so that the higher number always corresponds to the higher probability. In consequence of our fundamental

assumption our rules always imply the existence of a definite probability-number. The rules, as we stated before, are conventions and not hypotheses; for if the probability-number assigned by our rules is x, any function of x that always increases with x would satisfy the fundamental assumption. But the choice that we have made seems to be far the most convenient. Henceforth we shall have no need to speak of probabilities apart from their associated numbers, and when we speak of the probability of a proposition on given data we shall mean the number associated with the probability by our rules.

2·3. We now introduce the notation $P(p \mid q)$ for the probability of the proposition p on the data q*. It may be read "the probability of p given q". We also adopt the following notations from mathematical logic.

$\sim p$ means the contradictory of p, that is, the statement that p is untrue. It is read "not p".

$p \vee q$ means the disjunction of p and q, that is, the proposition that at least one of p and q is true. It applies whether p and q are consistent with each other or not. It is read "p or q".

* W. E. Johnson and J. M. Keynes use the notation p/q for the probability of p given q. The disadvantage of this notation is that the oblique stroke is a recognized device for printing fractions. As actual fractions will often occur explicitly in this work it seems desirable to avoid the confusion in reading that would arise from a similarity in notation.

In the earlier papers by Wrinch and Jeffreys† the notation $P(p : q)$ was used. The use of P calls attention directly to the fact that the number is a probability-number, and therefore to the fact that the elements within the bracket are propositions, and avoids complexity when the product of several probabilities has to be written. But the colon has the drawback that in the notation of mathematical logic it is often wanted for a bracket. The vertical stroke also has a meaning in mathematical logic, but there is no likelihood of confusion.

† "On Some Aspects of the Theory of Probability", *Phil. Mag.* **38**, 1919, 715–731. "On Certain Fundamental Principles of Scientific Inquiry", *Phil. Mag.* **42**, 1921, 369–390; (second paper), *Phil. Mag.* **45**, 1923, 368–374. "The Theory of Mensuration", *Phil. Mag.* **46**, 1923, 1–22. "The Relation between Geometry and Einstein's Theory of Gravitation", *Nature*, **106**, 1921, 806–809.

$p \cdot q$ means the proposition that p and q are both true. It is called the joint assertion of p and q, and is read "p and q".

These notations may be combined. Thus $\sim(p \cdot q)$ means the proposition that p and q are not both true, and therefore is equivalent to $\sim p \vee \sim q$.

Evidently

$$P(p \mid p) = 1; \quad P(\sim p \mid p) = 0. \tag{1}$$

2·31. Now suppose we have a set of data h. Then the following four propositions are mutually exclusive: $p \cdot q$, $\sim p \cdot q$, $p \cdot \sim q$, $\sim p \cdot \sim q$. By our original rule the probability that one of $p \cdot q$ and $p \cdot \sim q$ is true is the sum of their probabilities separately. But one of $p \cdot q$ and $p \cdot \sim q$ is true if, and only if, p is true.

Hence

$$P(p \mid h) = P(p \cdot q \mid h) + P(p \cdot \sim q \mid h). \tag{1}$$

Similarly

$$P(q \mid h) = P(p \cdot q \mid h) + P(\sim p \cdot q \mid h). \tag{2}$$

By addition

$$P(p \mid h) + P(q \mid h) = 2P(p \cdot q \mid h) + P(p \cdot \sim q \mid h) + P(\sim p \cdot q \mid h). \tag{3}$$

But the disjunction $p \vee q$ is true if and only if one of $p \cdot q$, $\sim p \cdot q$, and $p \cdot \sim q$ is true. Hence

$$P(p \vee q \mid h) = P(p \cdot q \mid h) + P(p \cdot \sim q \mid h) + P(\sim p \cdot q \mid h), \tag{4}$$

and therefore, comparing the last two equations, we have

$$P(p \mid h) + P(q \mid h) = P(p \vee q \mid h) + P(p \cdot q \mid h). \tag{5}$$

2·32. Consider next a class of n propositions, of which we know that one and only one is true, and any one is as probable as any other. Then if any m of them are selected the probability that one of these m is true is m/n. Let q denote the

proposition that one of these m is true, and h the data we had initially. Then

$$P(q \mid h) = m/n. \tag{1}$$

Consider another class of the original propositions and let p denote the proposition that some member of this class is true. Then the proposition that p and q are both true is the proposition that some proposition in the common part of the two classes is true. Let the number of propositions in the common part be l. Then

$$\begin{aligned} P(p \cdot q \mid h) &= l/n \\ &= (l/m)\,(m/n). \end{aligned} \tag{2}$$

Now consider $P(p \mid q \cdot h)$, the probability that p is true given h and q. h and q are both true if the true proposition is included in the class of number m. p is true, given q and h, if the true proposition is one of the common part, of number l, given that it is one of the class of number m. Hence

$$P(p \mid q \cdot h) = l/m, \tag{3}$$

and finally

$$P(p \cdot q \mid h) = P(p \mid q \cdot h)\, P(q \mid h). \tag{4}$$

This proposition is of capital importance. We have proved it for cases where p and q are expressible as disjunctions of equally probable and mutually exclusive alternatives. It cannot be proved in general without some further assumption. If $P(p \cdot q \mid h)$ was a function of $P(p \mid q \cdot h)$ and $P(q \mid h)$, different from their product, then we could choose l, m, and n so as to make the theorem untrue in some of the cases where we have proved it true; but we cannot absolutely exclude the possibility of another variable entering into the equation and producing exceptions to the rule (4) when the probabilities are not R-probabilities. It does not seem worth while, however, to consider such a possibility at present. It will be assumed without further discussion that (4) holds in general.

2·33. We have also by symmetry

$$P(p \cdot q \mid h) = P(q \mid p \cdot h)\, P(p \mid h), \qquad (5)$$

and therefore

$$P(p \mid q \cdot h) = \frac{P(q \mid p \cdot h)}{P(q \mid h)} P(p \mid h). \qquad (6)$$

This theorem yields as an immediate consequence the principle of *inverse probability*. Suppose that q is a logical consequence of p and h, so that $P(q \mid p \cdot h) = 1$, and suppose further that q has been verified. Then $P(p \mid h)$ is the prior probability of p before the verification and $P(p \mid q \cdot h)$ the posterior probability after the verification. Then our result is that the posterior probability of p is the prior probability of p divided by the prior probability of the consequence. The more remarkable the consequence, then, the greater the increase produced by its verification on the probability of the hypothesis under test.

2·34. Again, suppose that $p_1, p_2, \ldots p_n$ are a number of mutually exclusive hypotheses such that one of them must be true. Then for each we have a relation of the form (6), and therefore

$$\frac{P(p_1 \mid q \cdot h)}{P(q \mid p_1 \cdot h)\, P(p_1 \mid h)} = \frac{P(p_2 \mid q \cdot h)}{P(q \mid p_2 \cdot h)\, P(p_2 \mid h)} = \ldots = \frac{P(p_n \mid q \cdot h)}{P(q \mid p_n \cdot h)\, P(p_n \mid h)}. \qquad (1)$$

But $$\sum_{r=1}^{n} P(p_r \mid q \cdot h) = P(p_1 \vee p_2 \ldots \vee p_n \mid q \cdot h), \qquad (2)$$

since the p's are mutually exclusive

$$= 1, \qquad (3)$$

since it is known that one of the p's is true. Hence each of the fractions in (1) is equal to

$$\frac{1}{\sum_{r=1}^{n} P(q \mid p_r \cdot h)\, P(p_r \mid h)}.$$

Therefore

$$P(p_r \mid q \cdot h) = \frac{P(q \mid p_r \cdot h)\, P(p_r \mid h)}{\sum_{r=1}^{n} P(q \mid p_r \cdot h)\, P(p_r \mid h)}. \qquad (4)$$

This theorem* is to the theory of probability what Pythagoras's theorem is to geometry.

2·341. It follows at once that $P(p_r \mid q \cdot h)$ can hardly ever be unity; for in the fraction on the right the denominator is the sum of the numerator and a number of other positive terms. But if q has a small probability on all the hypotheses except one, p_1 say, and a large probability on that one, and the prior probabilities of the hypotheses are comparable, then the posterior probability of p_1 may approach unity. This is the type of inference known as a *crucial test*.

2·342. Again, suppose that p_1 implies $\sim q$, so that

$$P(q \mid p_1 \cdot h) = 0,$$

and that nevertheless q is verified. Then (4) shows that

$$P(p_1 \mid q \cdot h) = 0.$$

This explains how the failure of a crucial test may reduce a previously plausible hypothesis to impossibility.

2·343. It may happen that the probability of q is the same on all the hypotheses under discussion; that is, that

$$P(q \mid p_r \cdot h)$$

is the same for all values of r. Then

$$P(p_r \mid q \cdot h) = \frac{P(p_r \mid h)}{\sum_{r=1}^{n} P(p_r \mid h)}. \qquad (1)$$

But the p_r's are known to be mutually exclusive, and one of them is true. Hence

$$\sum_{r=1}^{n} P(p_r \mid h) = 1,$$

and

$$P(p_r \mid q \cdot h) = P(p_r \mid h). \qquad (2)$$

* Bayes, *Phil. Trans.* 53, 1763, 376–398.

Thus for each hypothesis the posterior probability is equal to the prior probability, and the test does nothing to help us to decide between the hypotheses. This is the case of *irrelevance.*

2·344. On the other hand suppose that q is a logical consequence of h alone. Then $P(q \mid h)$ and $P(q \mid p \,.\, h)$ are both unity, and

$$P(p \mid q \,.\, h) = P(p \mid h). \qquad (1)$$

If for instance h consists of the primitive propositions of logic and mathematics, and q is any demonstrated proposition of pure mathematics, then q can be included in the data without affecting any probability.

It may be mentioned that the case where q is implied by p and contradicted by h cannot arise; for $P(q \mid p \,.\, h)$ depends on the possibility of p and h being data at the same time, and this cannot happen if one implies a consequence contradicted by the other, for then they would be inconsistent.

2·4. In all estimates of posterior probability by means of the theorem of 2·34, the prior probabilities of the hypotheses appear explicitly. The theorem does not therefore give definite answers unless these prior probabilities are known; and here we come upon the greatest stumbling-block in the theory of probability.

How do we assess the probability of a proposition before we have any means of knowing whether it is true or false? It has often been said that assessing a probability implies some knowledge, and that therefore we cannot assign a probability when we are in complete ignorance. This opinion must be directly contradicted. Complete ignorance *is* a state of knowledge, just as much as a statement that a vessel is empty is a statement of how much there is in it, and the probabilities assigned upon it are perfectly definite. If we have no means of choosing between alternatives, the probabilities attached to those alternatives are equal*. If there are n alternatives

* This is usually known as the Principle of Sufficient Reason.

just one of which must be true, the prior probability of each is $1/n$.

The issue is fundamental. Either we can learn from experience or we cannot. The ability to learn from experience demands the concept of probability in relation to varying data, and the recognition of the meanings of *more probable than* and *less probable than*. Using only rules based on these concepts, we have shown how probabilities can be assessed. We must either accept the results or reject the fundamental principle and say that it is impossible to learn from experience. Whatever subject we take up, we start from ignorance and build up knowledge by means of experience. Everybody but a few philosophers recognizes the general validity of the process; and even the philosophers that say that they reject it show by their actions that their rejection is purely academic. Put the most sceptical philosopher in the situation described at the beginning of this chapter, and he will behave just like anybody else and probably express the same doubts.

But we have still not stated the method completely. Imagine a new-born baby to have seen only two objects, one blue and one yellow. Another object is to be introduced from outside. What is the probability that that object will be blue? If the alternatives are that it must be either blue or yellow the correct probability is $\frac{1}{2}$. But this is the probability on the datum that only two colours are possible. If the next object introduced proves to be pink this datum is proved wrong, and the fact that the probability was correctly assigned for it ceases to be of practical interest. This is a situation that we must accept in practice; we are often in situations where we cannot foresee every possible alternative, and allowance for the possibility of unforeseen alternatives must be made.

The issue can be stated in two simple ways.

1. The new object will be either blue, yellow, or some other colour. If we treat these as three equivalent alternatives the probability of each is $\frac{1}{3}$.

2. The new object will either have a colour known already or a new one. If we treat these as two equivalent alternatives the probability of each is $\frac{1}{2}$ and the probability that the next object will be blue is $\frac{1}{4}$.

Neither suggestion is quite satisfactory. "Some other colour" implies a choice among all possible other colours, which may be 0 to infinity in number, and it is not obvious that it can be treated as on an equivalent footing with one definite colour*. Nor is it obvious that, when the very existence of any other colour is problematic, some other colour is as likely to turn up as one of those already known.

The second suggestion is obviously wrong. If it were correct to treat the known and the unknown as equivalent alternatives, we could never, however many colours had been observed, have any additional confidence that the next one would have a colour already known. It therefore contradicts our fundamental postulate, that it is possible to learn from experience. What may be the correct answer will be indicated in the next chapter.

2·5. But the usual difficulty in assessing a prior probability at the beginning of an investigation does not arise from ignorance. The customary obstacle is too much knowledge. The statement that a probability number exists in every state of knowledge is not the same as the statement that we know what it is. The point may be illustrated from the purest of pure mathematics, the theory of numbers. How many prime numbers are there less than a billion? There *is* a number of such numbers; authorities on the subject can even say approximately what it is†; but just exactly how many prime numbers there are under a billion is known to nobody. It could be found out by trial, given sufficient time; but nobody has yet had time to do it. This is our usual situation in assessing a probability. The number of relevant facts is great,

* Keynes has, effectively, made this point. Cf. *Treatise on Probability*, 1921, 60.

† $10^{12}/12 \log_e 10$.

and their bearing on the probability of the proposition under discussion is difficult to evaluate precisely, though it may be easy in general terms. In actual life we simply do not take the trouble to evaluate the probability; we have not the time, for nobody can remember at once or enumerate all the relevant data at his disposal. If our traveller at the beginning of this chapter stopped to evaluate the probability accurately at any stage he would *certainly* miss his train.

The actual situation is therefore that the prior probabilities enter into our formulae, but we do not know their values, and they always affect the posterior probabilities. If this were not true newspapers would employ expert calculators of probabilities instead of unreliable turf tipsters. But in scientific work, though we can never make the posterior probability completely determinate, we can make it so near zero or unity as to amount to practical certainty or impossibility for *all* ordinary values of the prior probability. This is done by repeated verification and crucial tests. We do not know the prior probability of a scientific law when we begin an investigation of whether it is true; we swamp the prior probability by the number and variety of the verifications. The scientific man might, if he took enough trouble, evaluate the prior probability accurately; but in practice he is not interested in the accurate evaluation of a moderate probability. He prefers to obtain such additional information as will make the posterior probability approach impossibility or certainty whatever the prior probability may have been; and when that is done he no longer needs to evaluate the prior probability. Nevertheless it leaves its traces. The practical certainty or impossibility of an inference from abundant experience is not the same thing as absolute certainty or impossibility, which can come only from direct sensation or *a priori* knowledge.

CHAPTER III

SAMPLING

Little drops of water,
Little grains of sand
Make the mighty ocean
And the glorious land.

JULIA CARNEY

3·1. We are now in a position to discuss one of the most important applications of the theory of probability, the theory of sampling. The first problem is as follows: There are n objects with a defining property a. Of these, r have a further property b. We select at random m of the objects. What is the probability that l of these will have the property b?

We need a definition of what we mean by *at random*. We mean that every possible selection of m objects from the original n is equally probable. The total number of ways of selecting m things from a set of n is denoted by ${}^nC_m$. It is called the number of combinations of n things taken m at a time, and it is shown in works on algebra to be equal to $\frac{n!}{m!\,(n-m)!}$, where $n!$ means the product of all the integers from 1 up to n. There are r accessible objects with the property b. We can select l of these in ${}^rC_l$ ways. The other $n-r$ objects have not the property b; and if the sample of m objects contains l with the property b it must contain $m-l$ without it. Hence in our sampling we choose $m-l$ objects from a class of $n-r$, and this can be done in ${}^{n-r}C_{m-l}$ ways. But any selection of l objects with the property b is consistent with any selection of $m-l$ objects without it, and therefore the total number of ways of selecting m things so that l of them will have the property b is ${}^rC_l \,.\, {}^{n-r}C_{m-l}$. But by hypothesis all the possible ${}^nC_m$ selections are equally probable and mutually exclusive, and one of them is certain to be

made. Hence the probability that any particular one will be made is $1/{}^nC_m$, and the probability that we shall make some one of a set of number ${}^rC_l\,{}^{n-r}C_{m-l}$ is

$$g(l) = \frac{{}^rC_l\,{}^{n-r}C_{m-l}}{{}^nC_m}. \tag{1}$$

Since the set of number m must contain either 1, 2, ... or m things with the property b, the sum of the probabilities of the various values of l must be unity. Hence

$$\sum_{l=1}^{m} {}^rC_l\,{}^{n-r}C_{m-l} = {}^nC_m. \tag{2}$$

It is easily proved directly by algebra that this is the case.

We have

$$g(l) = \frac{r!}{l!\,(r-l)!} \cdot \frac{(n-r)!}{(m-l)!\,(n-r-m+l)!} \cdot \frac{m!\,(n-m)!}{n!}, \tag{3}$$

and therefore

$$\frac{g(l+1)}{g(l)} = \frac{r-l}{l+1} \cdot \frac{m-l}{n-r-m+l+1}, \tag{4}$$

which is greater or less than 1 according as

$$l+1 \text{ is less or greater than } \frac{(r+1)(m+1)}{n+2}. \tag{5}$$

The last quantity differs from rm/n by

$$\frac{r+m+1}{n+2} - \frac{2rm}{n(n+2)}, \tag{6}$$

which is always less than unity. It follows that changing l to $l+1$ will increase $g(l)$ if l is less than mr/n, save for a fraction, and will decrease it if l is greater than that value. Hence the most probable value of l is the integer nearest to mr/n; that is, the most probable sample is the *fair sample*, such that the number of things with the property b in the sample bears the same ratio to the total number of the sample as the number of things with the property b in the whole class bears to the number of the whole class.

For moderate values of n, m, r, and l the exact solution (1) tells us all we need. But if we have a large class to begin with, and extract from it a large sample, it can be proved (Lemma II) that the sum of the values of $g(l)$ for values of l between $mr/n + p_1$ and $mr/n + p_2$ is very nearly $\frac{1}{2}(\operatorname{erf}\xi_1 - \operatorname{erf}\xi_2)$, where

$$\xi = \{2r(n-r)m(n-m)/n^3\}^{-\frac{1}{2}}p. \qquad (7)$$

$\operatorname{erf}\xi$ vanishes for $\xi = 0$, but rapidly approaches $+1$ for moderate positive values of ξ, and -1 for negative values*. If then ξ_1 is a moderate positive number and ξ_2 a moderate negative one, the sum of the values of $g(l)$ corresponding to intermediate values of ξ will be nearly unity. The corresponding range in l is such that l varies in it by a moderate multiple of

$$H = \{2r(n-r)m(n-m)/n^3\}^{\frac{1}{2}}. \qquad (8)$$

Then H may be taken as a measure of the range of values of l that are probable. We notice that if r and $n-r$ are both moderate fractions of n, so that the original class was not overwhelmingly b or not-b, and if the sample is only a small fraction of the original set, so that m/n is small, then

$$2r(n-r)(n-m)/n^3, \qquad (9)$$

which is at most $\frac{1}{2}$, will be comparable with its maximum value. Then H is about $(\frac{1}{2}m)^{\frac{1}{2}}$. The range of probable deviation from a fair sample is of the order of $m^{\frac{1}{2}}$, where m is the number of the sample. In general the probability that the number of b's in the sample is between $l_0 \pm H$, where l_0 is the most probable value, is 0·843; the probability that it

* The following table will illustrate the point.

ξ	$\operatorname{erf}\xi$	ξ	$\operatorname{erf}\xi$
0	0	1·0	0·84
0·2	0·22	1·4	0·95
0·4	0·42	1·8	0·989
0·6	0·60	2·2	0·998
0·8	0·74	∞	1·000

For negative values $\operatorname{erf}(-\xi) = -\operatorname{erf}\xi$.

is between $l_0 \pm 2H$ is 0·995; and the probability that it is between $l_0 \pm 3H$ is indistinguishable from unity.

As a specimen of the numerical results, consider the case where the original class is equally divided, so that $r = \frac{1}{2}n$ and $H = (\frac{1}{2}m)^{\frac{1}{2}}$. Take $m = 100$. Then $l_0 = 50$, $H = 7$, and the probability is 0·995 that l will lie between 36 and 64. If instead we take a sample of number 10,000, $l_0 = 5000$, $H = 71$, and the probability is 0·995 that l will lie between 4858 and 5142. We notice the large size of the sample that has to be taken to establish a high probability that the sample will be fair within 1 per cent. of its total number.

3·2. In the above discussion we have supposed the composition of the whole class known, and we have determined the probabilities of different compositions of the sample. The usual problem of sampling is the inverse one: given the composition of the sample, what inferences can we draw about the composition of the whole class? We make use of formula 2·34 (4). Here let p_r denote the proposition that there are just r things in the original class with the property b; then $P(p_r \mid h)$ is the prior probability that this value of r is correct. Let us denote it by $f(r)$. The verified proposition q is here the fact that the known sample consists of l things with the property b and $m - l$ things without it. Then $P(q \mid p_r \cdot h)$ is the probability that a sample, m in number, drawn from a class known to consist of r things with the property and $n - r$ without it, would contain just l things with the property. This is the function we called $g(l)$, namely,

$$g(l) = \frac{{}^{r}C_{l}\ {}^{n-r}C_{m-l}}{{}^{n}C_{m}}. \qquad (1)$$

Since l is now to be kept constant while r varies, we shall now call this function $h(r)$. Now applying 2·34 (4) we have

$$P(p_r \mid q \cdot h) = \frac{f(r)\, h(r)}{\sum_{r=1}^{n} f(r)\, h(r)}. \qquad (2)$$

As usual we can make no further progress without some knowledge of the form of the prior probability $f(r)$. If there is no previous reason to think one value of r more likely than any other, $f(r)$ is the same for all values of r. In that case the posterior probability reduces to

$$P(p_r \mid q \,.\, h) = \frac{{}^rC_l \, {}^{n-r}C_{m-l}}{\sum_{r=0}^{n} {}^rC_l \, {}^{n-r}C_{m-l}}. \qquad (3)$$

It is easy to show that this has its greatest value when r/l is as near as possible to n/m; that is, the most probable value of r is found by supposing that the known sample is a fair one.

Suppose that we wish to know the probability that the next object examined will have the property b. Denote this proposition by q'. Then the probability required is

$$P(q' \mid q \,.\, h).$$

Since one and only one of the propositions p_r is true

$$\begin{aligned} P(q' \mid q \,.\, h) &= \sum_r P(q' \,.\, p_r \mid q \,.\, h) \\ &= \sum_r P(p_r \mid q \,.\, h)\, P(q' \mid p_r \,.\, h). \qquad (4) \end{aligned}$$

To evaluate $P(q' \mid p_r \,.\, h)$ we must suppose a definite value of r chosen. l things with the property b have already been removed, and therefore $r - l$ remain. The total number of things left to choose from is $n - m$. Hence the probability of picking a thing with the property b at the next attempt is

$$P(q' \mid p_r \,.\, h) = \frac{r-l}{n-m}.$$

Then

$$P(q' \mid q \,.\, h) = \frac{\Sigma_r \frac{r-l}{n-m} \, {}^rC_l \, {}^{n-r}C_{m-l}}{\Sigma_r \, {}^rC_l \, {}^{n-r}C_{m-l}}, \qquad (5)$$

which is equal, by Lemma III, to

$$\frac{(l+1)\,(n+1)!}{(n-m)!\,(m+2)!} \Big/ \frac{(n+1)!}{(m+1)!\,(n-m)!} = \frac{l+1}{m+2}. \qquad (6)$$

We notice that the probability of drawing another thing with the property b at the next attempt depends wholly on the composition of the sample already drawn, and not on n. If m is large and l equal to m, this probability approaches unity, but never quite reaches it.

Consider next the probability that the whole of the class may have the property b. For the possibility to arise it is obvious that all the known instances must be b's; that is, l must equal m. In this case $r = n$, and

$$P(p_n \mid q \cdot h) = \frac{n!}{m!(n-m)!}\frac{0!}{0!0!}\Big/\frac{(n+1)!}{(m+1)!(n-m)!}$$

$$= \frac{m+1}{n+1}. \qquad (7)$$

This approaches unity only when m is nearly n, that is, when nearly the whole of the class has been examined. It appears that pure sampling methods will never establish a high probability for the proposition that the whole of the set is of one type.

3·3. The above analysis, which is due to Laplace, has been repeatedly attacked. It obviously depends fundamentally on the form of the function $f(r)$, which is taken constant by Laplace, and represents the prior probability that a value of r is correct. But if our data included the proposition that the number of balls with the property b in the bag is just s, say, the prior probability $f(r)$ is 1 for $r = s$ and zero for all other values of r. Referring back to 3·2 (2) we see that the posterior probability of a given r is also 1 for $r = s$ and zero for all other values of r, and is entirely unaffected by the composition of the sample—as we should of course expect. Also the objects unexamined are $n - m$ in number, and include $s - l$ with the property b. Hence the probability that the next one examined has the property b is $\frac{s-l}{n-m}$, and obviously decreases as l increases. We should of course expect this; the more b's have

been examined the less likely are we to find one among the unexamined objects. But it is qualitatively different from the result of Laplace's theory, which indicates that the probability of a b at the next trial increases steadily with the number of b's in the sample. Obviously if we take $l = s$ the whole of the b's have already been removed and the probability that one will be found at the next trial is zero.

3·31. The form of the prior probability is therefore a matter of great importance. Laplace's theory has often been criticized on the ground that there is no reason to suppose that $f(r)$ is constant, and therefore that the theory rests on no foundations whatever. But this criticism misses the whole point of the theory. It is an instance of that already discussed in 2·4. Either we have reasons to prefer one value of r to another or we have not. In the latter case $f(r)$ is definitely constant and Laplace's theory is correct. In the former case Laplace's theory is simply inapplicable. The theory is in fact right when we have no previous knowledge of the composition of the class, but becomes inapplicable when we have relevant knowledge before we take the sample. The introduction of the function $f(r)$ makes it possible to allow for previous knowledge.

Though Laplace's theory is correct in the circumstances specified, the cases where it is not applicable are very numerous and important. Suppose for instance that there are n balls in a well-shaken bag, that they are known to be all of the same size, and that we have been told that one of them is a cricket ball and red. Then there is a strong prior probability that all are cricket balls and red. The only likely alternative is that hockey balls, which are of the same size but white, may be mixed with them, and we know that appliances for different games are usually kept separate. If then r is the number of red balls in the bag, $f(r)$ is very large for $r = n$ and small for all other values.

On the other hand if we have merely observed by feeling

that the balls are about the size and weight of hockey or cricket balls, then $f(r)$ is large for $r = 0$ and $r = n$ and small for intermediate values. The extraction of a single ball then establishes practical certainty that all the balls are cricket balls or all hockey balls, as the case may be*.

Again, suppose that the balls are known to be tennis balls and to be awaiting use in a match. If l of them have been examined and found white, the probability that the next will be white, on Laplace's theory, is $(l+1)/(l+2)$. If $l = 2$ this is $\frac{3}{4}$. But the actual probability in these circumstances is nearly unity, since one knows by previous experience that only new and white balls would be used in a match. If on the other hand the balls belong to an ordinary player's set towards the end of the summer there is a considerable prior probability that most of them will be green, and this affects the posterior probability in the opposite direction. In all these cases the departure of $f(r)$ from constancy materially affects the posterior probability. In addition we have the general knowledge that like things tend to be associated, as in the case of the cricket and hockey balls. Allowing for this we should expect $f(r)$ to be larger for r small and r nearly equal to n than for intermediate values. In some cases where $f(r)$ is not uniform it can be evaluated and the posterior probability can be found completely. But the determination of $f(r)$, when it is not constant, is usually troublesome, and we shall show that it is also often unnecessary.

3·4. Suppose now that the original class had number n, where n is very large, and that the number of the sample, m, is also large. The posterior probability that the whole class contained r things with the property b is

$$\frac{f(r)\,h(r)}{\sum_{r=1}^{n} f(r)\,h(r)},$$

* See also later, 10·1.

where, by Lemma II,

$$h(r) = g(l) = \left(\frac{n}{2\pi xyr(n-r)}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{np^2}{r(n-r)xy}\right); \quad (1)$$

$$x = m/n; \quad y = (n-m)/n; \quad (2)$$

$$l = rm/n + p. \quad (3)$$

We are now treating l as known and r as variable; the problem is the inverse of that of 3·1. The exponential is greatest when $p = 0$, that is, when

$$r = nl/m = r_0, \quad (4)$$

say. Put

$$r - r_0 = n\theta, \quad (5)$$

so that θ measures the departure of the composition of the whole class from that of the sample. Then

$$p = l - rm/n = (m/n)(r_0 - r) \quad (6)$$

$$= m\theta. \quad (7)$$

The index of the exponential is therefore, neglecting θ^3,

$$-\frac{1}{2}\frac{n^3m^2\theta^2}{r_0(n-r_0)m(n-m)} = -\frac{1}{2}\frac{nm^3\theta^2}{l(m-l)(n-m)}$$

$$= -h^2\theta^2 = -\xi^2, \quad (8)$$

say. When m is large absolutely, but small compared with n, h^2 is greater than $2m$, becoming equal to this minimum when $l = \frac{1}{2}m$. As in the case of direct sampling, therefore, the exponential factor is insignificant except within a range of values of θ comparable with $m^{-\frac{1}{2}}$. In these conditions we may ignore the variation of r in the factor outside the exponential in (1), and simply treat $h(r)$ as proportional to the exponential. The probability that the true value of r lies within a given range is then proportional to the sum of the values of

$$f(r)\,h(r)$$

for values of r within that range. When m is large, $h(r)$ is negligible except when θ, which measures the departure of the sample from fairness, is of order $m^{-\frac{1}{2}}$. The whole range of possible variation of r makes θ vary by unity. In these conditions, if $f(r)$ is a function of such types as usually arise,

appreciable contributions to $\Sigma f(r)\, h(r)$ arise only from the range of values that make the exponential moderate, and within this range $f(r)$ will not vary greatly. In ordinary cases $f(r)$ may be considerable for r small and r nearly equal to n, and may have one minimum between. Then when m is great we can treat $f(r)$ as constant within the range that matters, and it cancels from the numerator and denominator of the posterior probability. The sum may now be replaced by an integral, and the probability that r lies between $r_0 + n\theta_1$ and $r_0 + n\theta_2$ is proportional to $\int_{\theta_1}^{\theta_2} \exp(-h^2\theta^2)\, d\theta$ or $\int_{\xi_1}^{\xi_2} e^{-\xi^2}\, d\xi$. Hence the probability that r lies between $r_0 + n\theta_1$ and $r_0 + n\theta_2$ is

$$\int_{\xi_1}^{\xi_2} e^{-\xi^2}\, d\xi \Big/ \int_{-\infty}^{\infty} e^{-\xi^2}\, d\xi = \tfrac{1}{2}(\operatorname{erf}\xi_2 - \operatorname{erf}\xi_1)$$
$$= \tfrac{1}{2}(\operatorname{erf} h\theta_2 - \operatorname{erf} h\theta_1). \qquad (9)$$

This result shows that, except in cases so remarkable that they must be easily recognized if they arise, the actual variation of the prior probability with r is not important provided that the sample is large. This is the real reason why it is unnecessary in most cases to evaluate the prior probability. Within ordinary limits its effect on the answer is negligible. In fact the range of values of θ such that the truth is practically certain to lie within it is of order $m^{-\frac{1}{2}}$. To make this of order 1 per cent. we need a sample of number 10,000 or so; while the only important values of $f(r)$ come from values of r within a range of order 1 per cent. of the whole possible range, and in such a small range we cannot expect the variation of $f(r)$ to matter. In fact the large sample is necessary in any case, to ensure fairness; and when so large a sample has been taken that fairness is assured the results will in any case be the same as if the prior probability was constant.

We can now sum up the position concerning the prior probability in the theory of sampling. There is no theoretical difficulty. The opinion that there is any fundamental objec-

tion to the notion of prior probability can be maintained only at the cost of rejecting the notion of probability, and with it the universally accepted opinion that it is possible to start from ignorance and gradually build up from experience methods of predicting the truth. There are practical difficulties in assessing the prior probability in many cases as they actually arise. This is not a situation to evade, but one to face. It could be dealt with in two ways: we may either evaluate the prior probability or swamp it. The former alternative is laborious and unnecessary; for in any case a large sample is needed to make it practically certain that the sample is a nearly fair one, and then the posterior probability of a given departure from fairness is almost the same whatever the prior probability may be. We do not evaluate the prior probability in practical sampling because we do not need to; we swamp it automatically when we take a sufficiently large sample. It is this principle that constitutes the theoretical justification of statistical methods.

3·5. We may now return to the problem of the baby that has seen only two objects, one blue and one yellow. What is the probability that the next object seen will be blue? The number of other colours in the world may be anything; but we can state the issue by considering "blue or yellow" as a single property, as opposed to "not blue or yellow". Then there are two known instances of the property "blue or yellow" and none of its absence. Thus in the theory of sampling $m = l = 2$, and the posterior probability that the next object seen will be blue or yellow is $(l + 1)/(m + 2)$ or $\frac{3}{4}$, whatever the total number of objects in the world may be. Since on the data blue and yellow are equally probable, the probabilities that the next object will be blue, yellow, or some other colour are respectively $\frac{3}{8}$, $\frac{3}{8}$, and $\frac{1}{4}$.

It would not be legitimate to proceed by treating blue and yellow as single independent alternatives. If for instance we treated blue as one alternative and "not blue" as the other,

then we should have $m = 2$, $l = 1$, and the probability of a blue object at the next trial will be $\frac{1}{2}$. Similarly the probability of a yellow one would be $\frac{1}{2}$, and taking the two together we should say that the probability that the next object will be blue or yellow is 1. This is absurd. The error is that in treating blue as a single alternative and applying Laplace's theory we suppose all numbers of blue things in the world equally probable *a priori*; similarly for yellow things. Thus we have made no allowance for the fact that it is impossible in the same circumstances for more than half the things in the world to be blue and more than half yellow; the prior probabilities of given numbers of blue and yellow things in the world are not independent.

The purpose of this trivial example is to illustrate the fact that allowance can actually be made in probability estimates for the possibility that an unforeseen alternative may arise.

CHAPTER IV

QUANTITATIVE LAWS

'Tis a lesson you should heed,
Try again;
If at first you don't succeed,
Try again;
Then your courage should appear,
For if you will persevere
You will conquer, never fear,
Try again.

WILLIAM EDWARD HICKSON

4·1. The majority of the laws of physics are of the form

$$y = f(x_1, x_2, x_3, \ldots x_n), \qquad (1)$$

where $y, x_1, x_2, x_3, \ldots x_n$ are quantities determined by measurement, and f is a known mathematical function. Such a law enables us to calculate y when the x's are known. These laws are established by repeated verification; it is found in numerous instances that the observed value of y agrees closely with that calculated from the law, and on the strength of this verification it is asserted that the law holds in general. Superficially the generalization bears a resemblance to that involved in Laplace's theory of sampling when all the objects examined have hitherto been of the same type, but we shall see that the differences are very great. For instance, we may say that the position of Jupiter, as calculated from the law of gravitation, has agreed with prediction every time it has been observed during a revolution. But at the best the number of verifications is finite. What is the probability that the position of Jupiter *always* agrees with the calculated value? We are here generalizing from a finite number of verifications to an infinite class of possible instances, and if we apply the ordinary rules of sampling we must say that the probability is infinitesimal. On the other hand any astronomer would say

that it is practically certain that the position of Jupiter always agrees with prediction—unless indeed he said it was absolutely certain.

4·2. The quantitative laws of physics therefore seem to be in a somewhat different position from the rules established by sampling, and further inquiry into their nature is desirable. Let us consider first a simple experiment. A solid of revolution can roll down an inclined plane, and its displacement is observed every fifth second after it starts from rest. If we denote the time by t and the displacement from the starting point by x, the observations are as follows:

t (seconds)	0	5	10	15	20	25	30
x (centimetres)	0	5	20	45	80	125	180

Then we can say that at all the instants of observation the displacement is connected with the time by the formula

$$5x = t^2. \qquad (2)$$

On the face of it this statement is a pure description of observed facts. The phenomenalist school of critics would say that it is nothing more; and many physicists think that they belong to this school. But the facts of observation would be fitted equally well if the displacement was really connected with the time by the formula

$$5x = t^2 + t\,(t-5)\,(t-10)\,(t-15)\,(t-20)\,(t-25)\,(t-30)\,f(t), \qquad (3)$$

where $f(t)$ may be any function whatever that is not infinite at $t = 0, 5, 10, \ldots 30$ seconds. The law (2) is indeed not the only description that fits the data; it is only one of an infinite number of laws that would fit the data equally well. Its special quality that distinguishes it from the other possible laws is its *simplicity*. In practice no physicist, looking at the above data, would hesitate to say that the law (2) is the correct way of expressing them. But different physicists would disagree about their reasons for adopting it. Some

would say that it is a matter of strict necessity. This is just false; there are an infinite number of other alternatives that might be adopted if we chose. We want to know why there is only one that we *would* choose. Others would say that the simplest law is chosen for the sake of convenience. But a simple test would show that this is not the real reason. Suppose we want to know where the body was 18 seconds from the start. According to the law (2) it would be 64·8 cm. from the starting point. But according to the more general description (3) it might be anywhere, according to the value of $f(t)$ for $t = 18$ seconds. But what would happen if we said this to a physicist? He would certainly say "Don't be silly". Suppose that we pressed him, and that as a result he was persuaded to repeat the experiment and found that the displacement 18 seconds from the start was 55 cm. He would still not abandon the *form* (2). He would do the whole experiment again in order to find out why x/t^2 had changed; and he would expect to find that in the new experiment the values of x/t^2 at different times were again all equal, and different from the value 0·2 cm./sec.^{2} which he found before. (Further, he *would* find this to be so; and he would probably attribute the change to an alteration in the slope of the plane. But that is not our present point.) In fact the physicist, having once found x proportional to t^2 for a wide range of values of t, feels a complete confidence that this rule holds for other values of t. This confidence could not exist if he had chosen the simple law merely because it was convenient. He must have chosen it because, of all the laws that would fit the data, the simplest is the most likely to be correct for other values of the variables.

Let us put the matter in another way. Some physicists would say that the law (2) is adopted because it is *observed* to be true. But this statement is merely a mathematical pun. What is observed is that for $t = 0, 5, 10, 15, 20, 25, 30$ seconds, $5x = t^2$. What is asserted is that for *all* values of t, $5x = t^2$. The former statement is merely a concise way of

rewriting the observations, a shorthand description. The latter is an inference from the finite number of actual observations to an infinite number of possible observations. To express both by saying simply "$5x = t^2$" is to use the same language to mean two different things. In the same way, the law (3) applied to the observed values is definitely true; but no physicist would apply it to the unobserved values. In fact the preference for the simple law enters the question *only* when the need for making inferences arises. Convenience of description has nothing to do with the matter, unless we choose to say that "$5x = t^2$ for other than observed values of t" is a description. That is another pun; to describe an observation that has been made obviously does not mean the same thing as to describe an observation that has not been made. The word *description* is here restricted to descriptions of observed events; other events are *inferred*, not *described*.

We have seen that if we have a set of possible general laws p_r and a verified consequence q whose probability on all the laws is the same,

$$P(p_r \mid q \cdot h) = \frac{P(p_r \mid h)}{\sum_{r=1}^{n} P(p_r \mid h)}.$$

In other words, $P(p_r \mid q \cdot h)/P(p_r \mid h)$ is the same for all the laws. Now in our case law (2) and all the laws (3) imply the observed facts. Hence their posterior probabilities are in the same ratios as the prior probabilities. The physicist's confidence in the generality of the simple law in comparison with complex ones that fit the observations equally well must therefore correspond to an overwhelmingly greater prior probability for the simple law. Here, then, we come upon the essence of the problem. The prior probability of a simple law is so great in comparison with those of complex ones that, from a physicist's point of view, the latter are not worth considering. It is this fundamental principle that accounts for the physicist's preference for the simple law.

The above argument is actually an understatement of the

situation. In the inclined plane experiment the observations would not, in fact, fit the simple law exactly. We might get a series of values like the following:

t (seconds)	0	5	10	15	20	25	30
x (centimetres)	0	5	19	44	81	124	178

These do not fit exactly the law $5x = t^2$, or any other simple square law. But it would be easy to find a polynomial of the form

$$x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 + a_6 t^6,$$

that would fit the observations exactly. Nevertheless the physicist would stick to the square law. His expressed reason would be interesting. It would be that *any* set of seven values whatever can be represented by an expression with seven adjustable constants. Consequently the expression so obtained tells us nothing with regard to the reliability of the determination. The very fact that the representation is of such generality that it can always be made to fit the data exactly is considered an argument against it, not for it. With regard to the original square law, he would say that the observed values never differ from the calculated ones by more than 1 cm., except for the last; this differs by 2 cm., but at the time the velocity is 12 cm./sec., and the difference could be accounted for by an error in timing of 0·17 second, while the observations were made only to 0·2 second. In fact he would say that the differences never exceed the admissible errors of observation, and that the agreement of the observations with the simple law is perfectly satisfactory.

Apart from the physicist's specified reasons, which we shall have occasion to consider later, we notice the outstanding fact about his decision. His predilection for the simple law is so strong that he will retain it, even when it does not fit the observations exactly, in spite of the existence of complex laws that do fit them exactly. Simplicity is a better guarantee of probability than accuracy of fit. The physicist would use the square law to predict the value of x for $t = 60$ seconds,

and would expect the result to be right within a few centimetres, provided the plane was long enough to permit the displacement required. He would, on the other hand, expect the polynomial of seven terms to give a seriously wrong answer when extrapolated to such an extent.

The actual behaviour of physicists in always choosing in practice the simplest law that fits the observed facts therefore corresponds exactly to what would be expected if they regarded the probability of making correct inferences as the chief determining factor in selecting a definite law out of an infinite number that would satisfy the observations, and if they considered the simplest law as having far the greatest prior probability. It is not explained by the reasons that are usually stated.

4·3. We may also consider the problem as one of pure theory of probability without considering the behaviour of the physicist. We return to the law

$$P(p \mid q \cdot h) = \frac{P(q \mid p \cdot h)}{P(q \mid h)} P(p \mid h). \qquad (1)$$

Suppose that p is the general law whose probability we are considering, and that $q_1, q_2, \ldots q_n$ are successive verified predictions from it. If q_2 is implied by p, it is also implied by p and q_1 together. Thus we have in turn

$$P(p \mid q_1 \cdot h) = \frac{P(p \mid h)}{P(q_1 \mid h)},$$

$$P(p \mid q_1 \cdot q_2 \cdot h) = \frac{P(p \mid q_1 \cdot h)}{P(q_2 \mid q_1 \cdot h)},$$

.....

$$P(p \mid q_1 \cdot q_2 \cdot \ldots q_n \cdot h) = \frac{P(p \mid q_1 \cdot q_2 \cdot \ldots q_{n-1} \cdot h)}{P(q_n \mid q_1 \cdot q_2 \cdot \ldots q_{n-1} \cdot h)}.$$

Thus each successive verification divides the probability of the law by the probability of the verification on the data already known. Now if all the numbers of the form

$$P(q_n \mid q_1 \cdot q_2 \cdot \ldots q_n \cdot h)$$

were less than some proper fraction r, and p had a finite probability at any stage of the investigation, then a sufficient number of further verifications would give p a probability greater than unity, which is impossible. Hence we have to choose between two alternatives:

(1) However often p may be verified, its probability on the data is never finitely different from zero.

(2) The probabilities of the verifications, given in each case the previous verifications, are not all less than r if r is less than unity; that is, when the number of verifications becomes large, the probability of the next tends to unity as a limit.

The first alternative plainly does not agree with ordinary belief. However sceptical one may be about a given law that is consistent with the known facts, one would consider its probability finite. The second alternative, on the other hand, agrees perfectly with our fundamental belief in the possibility of acquiring knowledge by experience. But it says nothing about the probability of the law itself, but only of verifications of it. It might apparently be possible to adopt the second alternative and still suppose the probability of the general law infinitesimal.

But the construction of a satisfactory theory on such a basis would require a branch of mathematics that does not exist. Let us see whether a theory of quantitative inference can be constructed on the hypothesis that all general laws have the same prior probability. Suppose the number of such laws to be m, and suppose that a number of experiments have been made to test them. Then the only survivors are those that imply the results of the experiments, which may be summed up in the proposition q. Each of them after the experiments has the probability $1/mP(q \mid h)$. Thus every surviving law has the same probability after the experiments. Also since an infinite number of laws satisfy any finite number of measures, an infinite number survive, and the posterior probability of each is infinitesimal. Now suppose another verification to be

attempted. An infinite number of results are possible, corresponding to the different laws, and each result can be obtained from an infinite number of laws. The probability of a given numerical result at the next trial is therefore the ratio of two infinite numbers; and nobody has yet succeeded in constructing a satisfactory mathematical theory of such ratios. Until it is done we shall say that it is impossible to construct a theory of quantitative inference on the hypothesis that all general laws have the same prior probability.

4·4. Our effort to avoid the assumption that general laws have finite probabilities has thus led nowhere. Let us now make this assumption and investigate its implications. The number of possible laws is certainly infinite. How can an infinite number of mutually inconsistent laws all have finite probabilities? The answer to this question is provided by mathematics. Consider the series

$$\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\ldots.$$

The number of terms in this series is infinite, but every term is finite, the sum of any number of terms is less than unity, and the sum tends to unity as we take an increasingly large number of terms from the start. The assumption we need is therefore that the prior probabilities of possible general laws are the terms of a convergent series whose sum to infinity is unity. We have been led to it purely from the assumption that it is possible to construct a theory of quantitative inference; if this can be done such an assumption about the prior probabilities of laws must be made. Further, we see now that it fits in perfectly with our discussion of the relation of simplicity to prior probability; all we need to add is that the simpler the law is, the earlier its probability occurs in the series. Simplicity is a property that is easily recognizable when it is present, and we say that the order of decreasing simplicity among laws is also the order of decreasing prior probability.

If we make this assumption we find that there must be a severe restriction on the laws that are admissible at all. The terms of an infinite series are $\aleph_0$ in number*, and according to our rule no law whose probability is not a term of the series can ever be established by experience. Hence the quantitative laws capable of being established are $\aleph_0$ in number, and our problem is to specify a set of laws, $\aleph_0$ in number, that will include all laws required, or likely to be required, in physics.

In one sense it might be said that the problem is trivial, since the number of known physical laws at any time is finite, and likely to remain so. Nevertheless there is a theoretical problem apart from the actual facts; we are concerned not only with what is true, but with what is possible—or rather with what it would be possible to establish.

It is plain that not all functions are admissible in laws; for the number of all functions is C^C, which is greater than $\aleph_0$. The same applies to the class of continuous functions. Even if we restrict the functions to be analytic, the number of such functions is still C, or $2^{\aleph_0}$, which is greater than $\aleph_0$; not all analytic functions can occur in physical laws. If a physical law contains one numerical constant capable of continuous variation, the number of possible values of that constant is C; if the coefficients in the expression of an analytic function in a power series are restricted to be rational fractions, the number of functions is still C. The class of all polynomials of degree less than some finite number, and with rational fractions or algebraic numbers as coefficients, has number $\aleph_0$, but it does not include trigonometrical functions, which do occur in physics, and therefore is not sufficiently general.

Transcendental functions such as exponential, trigonometric, and Bessel functions do occur in physics, but we may notice that they are hardly ever derived directly from

* $\aleph_0$ is the number of positive integers, and is the smallest infinite number. C is the number of values of any quantity capable of continuous variation; and in particular is the number of real numbers. See Appendix.

observation. They arise first in theoretical work, and it is not till afterwards that it is verified that they do satisfy the results of observation. In the theoretical work they arise as the solutions of differential equations of finite order and degree.

4·5. Consider then the possibility of defining a class of differential equations, $\aleph_0$ in number. Clearly no numerical coefficient in such a class may be capable of more than $\aleph_0$ values, otherwise the hypothesis would be vitiated at the start. But if each equation is restricted to be of finite order and degree, and each coefficient in it to be capable of $\aleph_0$ values at most, then the conditions are satisfied. (See later, Appendix.) The natural possibilities to consider for the coefficients are that they may be whole numbers or rational fractions. The latter alternative appears more general, but is not so in fact, for any equation with rational coefficients can be converted into one with integral coefficients by merely multiplying by the least common denominator. There is indeed a definite advantage in choosing the former alternative; for an equation involving only integers with no common factor is equivalent to no other equation with the same property, whereas an equation with fractional coefficients is equivalent to an indefinite number of others with fractional coefficients. Thus the use of fractional coefficients would permit ambiguities in the arrangement of the equations in order of decreasing simplicity, which are avoided by the restriction to integers. All our data are therefore consistent with the following general principle:

Every quantitative law can be expressed as a differential equation of finite order and degree, in which the numerical coefficients are integers.

In the arrangement of such equations so that they correspond one by one with the positive whole numbers, we should begin by rationalizing each equation if it already contains roots. Then we should group the equations so that those with equal values of the sum of the order, the degree, and the

absolute values of the coefficients, were classed together. The number in each group is finite. We should then arrange the groups according to increasing values of this sum, and adopt some convention regarding the arrangement of the equations in the same group. Thus the equations occurring early in the series would have low order and degree, and the numerical coefficients in them would be small integers. They would therefore be *simple*, as the term is generally understood. We may indeed give a precise definition of the *complexity* of an equation by saying that it is the sum of the order, the degree, and the absolute value of the coefficients. If the complexity is thus defined, it is a determinate mathematical problem to say how many differential equations have complexity less than or equal to n. But it is difficult. When n is large the number is certainly larger than 2^n; I have not obtained a closer estimate. This, however, is enough for present purposes. The series $\sum_{n=1}^{\infty} 1/n$ does not converge. Hence the total probability of the laws of complexity n must decrease faster than $1/n$, and any one individually must have a prior probability less than $2^{-n}/n$.

It may be objected that some of the arbitrary constants involved in the solutions of the differential equations of physics are capable of continuous variation within definite ranges, and that therefore the true number of solutions is C, and we are no further forward. The reply is that the differential form, not the integrated one, is the fundamental physical law. The arbitrary constants, so called by mathematicians, are not arbitrary in physics; they are determined by the boundary conditions, and it seems that these conditions, in their fundamental forms, involve no more arbitrariness than the differential equations.

It was also objected, when this suggestion was first made in a slightly different form, that the restriction to differential equations was inconsistent with the ideas about the quantum theory that prevailed in 1921. My own view at the time was

that the orbits in an atom, in a stationary state, are describable by differential equations of the usual type (this was then the current opinion) and that the quantum jumps, involving discontinuous changes of velocity, should be regarded as boundary conditions. But there have been many quantum theories since then. Those of Heisenberg and Dirac appear to have replaced both the ultimate differential equations and the conditions of the quantum jumps by finite difference equations; and there is no objection to supposing that the ultimate laws are finite difference equations, for these may equally well be restricted to a class $\aleph_0$ in number. On the other hand Schrödinger's theory makes a single differential equation account for everything, and is entirely consistent with the postulate as it stood.

4·51. I do not wish, therefore, to maintain that this form of the simplicity postulate is necessarily the final one. I do maintain, however, that a postulate restricting the number of admissible laws to $\aleph_0$ is necessary, and that the prior probabilities must decrease rapidly with decreasing simplicity. Modifications of the present form may be needed to admit such systems as Dirac's; also in the laws that appear in the general theory of relativity, and indeed in elasticity, the simplicity of the symmetry relations may compensate for the large numbers of terms in the equations. Meanwhile the present form will serve our purposes.

Our everyday ideas on this subject, as in most others, are a complicated system based in part on experience and in part on principles believed independently of experience. The latter we call *a priori*. To disentangle the latter we have to argue backwards, just as in logic the discovery of the primitive postulates was subsequent to a great development of mathematics by forward reasoning. By analysing the processes involved in our forward scientific reasoning we detect the fundamental postulate that it is possible to learn from experience. This is a primitive postulate, pre-

sumably on the frontiers between *a priori* and empirical knowledge. The status of the laws of probability and the simplicity postulate is that of inferences from this principle.

4·6. The variables in the differential or difference equations include the time and the co-ordinates of position. These are still generally believed capable of continuous variation. But these are not real variables, but apparent variables. When we assert, for instance, Laplace's equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0,$$

we are not implying a choice among an infinite set of laws in any one of which x, for instance, may have a value chosen from a continuous set of possibilities. We assert that for all values of x, y, z corresponding to points outside matter this differential equation is satisfied by the potential. In the language of mathematical logic this equation should be written as follows.

Whatever P, x, y, z, V may be, if V is the gravitation potential at P, a point outside matter with co-ordinates (x, y, z), then

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0.$$

When a symbol is given all its possible values, and the differential equation is asserted for all of them, the symbol is only an apparent variable. There is no objection to apparent variables being capable of continuous variation; what matters is the form of the law, not the actual values of the variables in particular verifications. Similarly Poisson's equation at points inside matter could be written as follows.

Whatever P, x, y, z, ρ, V may be, if V is the gravitation potential and ρ the density at P, a point with co-ordinates (x, y, z), there is a constant f such that

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = -4\pi f\rho.$$

There is no difficulty, similarly, about the fact that density and mass may appear in our equations and are apparently capable of continuous variation; for we assert the laws for all their values, and they are only apparent variables.

4·7. The question of the probability to be attached to a quantitative inference can now be dealt with. If p is the most probable law on the data at any stage, and q an additional experimental fact, we have

$$\frac{P(p \mid q \,.\, h)}{P(\sim p \mid q \,.\, h)} = \frac{P(q \mid p \,.\, h)}{P(q \mid \sim p \,.\, h)} \frac{P(p \mid h)}{P(\sim p \mid h)}.$$

By the hypothesis we have just made about the prior probabilities of laws, $P(p \mid h)/P(\sim p \mid h)$ is not very small. If q be implied by p, we have $P(q \mid p \,.\, h) = 1$, while if the contradictory of p gives no particular inference about the truth of q, $P(q \mid \sim p \,.\, h)$ may be very small, especially if q involves accurate measurement. Hence, even if p has not a very large probability already, a single verification of a consequence not predicted by its contrary may make

$$P(p \mid q \,.\, h)/P(\sim p \mid q \,.\, h)$$

enormous, and therefore the posterior probability of p is nearly 1. In such circumstances the probability of a further inference q_2 from the law is practically that of the law itself, since the second term in the equation

$$P(q_2 \mid q \,.\, h) = P(p \mid q \,.\, h)\, P(q_2 \mid p \,.\, q \,.\, h) + P(\sim p \mid q \,.\, h)\, P(q_2 \mid \sim p \,.\, q \,.\, h)$$

is the product of two small factors. In such inference, then, there is no advantage to be gained by proceeding directly from the data to the further inference rather than by way of the general law, as has sometimes been suggested.

It will be noticed that the argument in the last paragraph depends on the smallness of $P(q \mid \sim p \,.\, h)$. If, however, $\sim p$ involves a moderately probable law which also leads to q as a consequence, this probability will not be small, and the

probability of p after the verification will stand to that of this alternative law in almost the same ratio as before. A new crucial test will be required to decide the issue between them.

If we have to decide between different simple laws, the prior probabilities of which are in any case moderate, the high posterior probability of a law arises from its verification. If a law p_1 implies that the measure of a length will be between 15·7 and 15·8 cm., and this is found to be true, then there is no posterior probability for a law that said it would be 45·0 cm. and very little for one that said that it might be anything from zero to a metre. So long as two laws are not widely separated in the order of simplicity, the decision between them rests on the quantitative tests and not on the prior probability.

But when the laws are widely separated in the order of increasing complexity the prior probability is all-important, even when the known facts would fit either. An important case that has arisen in practice is that of small variations in the numerical constants in fundamental laws. Suppose that a law contains a numerical constant 2, and that we propose to alter this constant to $2_{\frac{1}{1000000}}$. Then in accordance with our principle that a law must be cleared of fractions before it is placed in the order of descending probability, this law will now have to be treated as if it contained numerical coefficients running into millions, and its position in the series will be millions, probably billions, of places later than before. Its prior probability is accordingly insignificant. In fact a small change in a numerical coefficient is not a trivial matter; from the point of view of the prior probability of the law it is the most drastic change that can be made. As an example, we may consider the inverse square law of force in electrostatics, in which the index has been shown experimentally to be -2 within $\frac{1}{21600}$. Then the only law within the admissible range that has an appreciable prior probability is the exact inverse square one, and it is unnecessary to consider any others. Similarly, we could discard at sight the suggestion that the

perihelion of Mercury could be explained if the attraction of the sun varied inversely as the 2·000,000,016 power of the distance instead of as the exact inverse square. The exiguous prior probability of such a law puts it beyond consideration, apart from the inconsistency with the observed motion of the moon's perigee that led to its abandonment. In fact the law established with a high probability by experience is not an approximation to the simple law, but the exact simple law itself. Consequently extrapolation over an indefinitely wide range can be carried out with the full probability of the law. This is the justification of the inferences concerning conditions at the centre of the earth or millions of years ago that form so large a part of geophysics and cosmogony.

The rapidity with which an exact quantitative law can be established depends, then, first on its being sufficiently simple to have a moderate prior probability, and second, on its power to make precise predictions that can be tested. Subject to these two conditions a theory of quantitative inference can be constructed that will fully explain the confidence that physicists show in their predictions.

CHAPTER V

ERRORS

A snapper-up of unconsidered trifles.
SHAKESPEARE, *A Winter's Tale*

5·1. We saw that when the physicist found that the displacement of his solid down the inclined plane varied nearly as the square of the time from the start he would adopt the exact square law as a statement of the facts, in spite of the existence of more complicated laws that would fit the observations exactly; and we have shown how this procedure can be justified on the basis of the low prior probability of complicated laws, which renders them unreliable for the purpose of inference. Nevertheless he might not allow the matter to end there. He might seek for explanations of the departures of the observed values from those calculated from the law. In some sense the square law is true; but the quantities that satisfy it are not quite the data of observation. The physicist would say that it was impossible to measure the time absolutely accurately, because the watch could not be read to less than a fifth of a second; there was also some possibility of inaccuracy in measuring the position of a moving object; the watch and the position of the solid were not observed at precisely the same instant, since some time would elapse in looking from one to the other; and possibly the slope of the plane was not exactly uniform. Having reduced his observations he would say with confidence that the acceleration of a body of the actual form rolling down a uniformly inclined plane was constant; this would be his general law for the experiment, which he could extend to bodies of different design and to planes with different slopes. He would on the other hand recognize that exact verification of the law would require conditions not realized in the actual

experiment, and that the departures from the law had some explanation.

The physicist's attitude to observations is not the naïf realism attributed to him by some philosophers, which would make every observation a perfect statement of a fact about the real world. It is essentially a critical realism. There is a belief that there are *true values* of the quantities that he sets out to measure, but it is not believed that the observed values are anything but an approximation to these true values, which are in the last resort unknowable. The differences between the true and observed values are called *errors*.

In practice, not knowing the true values, we compromise. When we have a number of observations of one or more variables, a simple law is found to fit them approximately. The case of a single measurement carried out several times may be brought under this head, the law involved being merely one of constancy with regard to the time. The law may involve some parameters not known already, and it will usually be impossible, however these parameters are chosen, to make the law fit all the observations exactly. But we can choose them so as to fit the observations as closely as possible, though it is largely a matter of convention what criterion we adopt to measure the closeness of the fit. When we have chosen one such criterion the parameters in the law and the values of the function are unique. We call these the *adopted values*. In general they will differ from the true values, but will be nearer to them than the observed values. The differences between the adopted and observed values are called *residuals*. The differences between the true and adopted values are the *errors of the adopted values*.

In general the procedure may be summed up as follows. The observed values are found; they exist because they are measured, and there is nothing more to be said. A simple law is found to fit them approximately. This is a statement of fact. Then by a conventional process we find adopted values close to the observed values that fit the law exactly.

So far as the convention is at our disposal the adopted values have some arbitrariness, but with a given convention they are unique. The adopted values therefore exist. The existence of the true values, however, is a postulate, the validity of which will have to be examined. We notice at present that the observed values are more fundamental in experience than the simple law, that the simple law is more fundamental than the adopted values, and that the whole process of finding the adopted values could be carried out equally well if there were no such things as true values.

Provisionally we shall assume that the true values exist, that the exact simple law refers to certain specifiable conditions, and that the errors arise from the fact that the actual conditions of the experiment differ to some extent from these ideal ones. The practical justification for this assumption is that it is actually found that the more closely these conditions are realized the more accurately the simple law fits the observations, though it never fits them exactly. The ideal conditions always reduce to the removal of unconsidered variables. Thus in the problem of the rolling solid we should construct the plane so as to have as nearly uniform a slope as possible, and we should substitute electrical recording devices to record the time and displacement simultaneously instead of relying on eye observations. The aim is to make the time the only independent variable, and thereby to remove variations of the displacement that may be due to variations in anything but the time. It is here that *causality* enters: if when y is kept constant, $x = f(t)$, and if when y varies x differs from $f(t)$, the changes in x are said to be *caused by* the changes in y. This is the practical definition of causality. In the actual experiment the errors are said to be caused by the unconsidered disturbing factors.

The most fundamental type of error is inaccuracy of measurement. Observed values are never capable of taking all values of a compact set; in making a measurement we read the instrument to the nearest multiple of a certain constant,

which we call the *step* of the instrument. Thus in measuring the position of a mark on a scale we may read to the nearest hundredth of a centimetre; in observing an instant of time we give it to, say, the nearest fifth of a second.

5·2. The possible observed values in the one case are multiples of a hundredth of a centimetre, in the other, of a fifth of a second. The true value is not in general a possible observed value, since the true values of most variable quantities vary continuously. Hence there is an error of observation equal to the difference between the true value and the nearest observable value, which may in an extreme case be half the step of the instrument. Such an error, for a given true value, is systematic; that is, it is always the same however often we repeat the measurement.

5·21. Suppose next that we wish to measure a length by means of a scale. Take the step of the measuring scale as the unit of length and suppose that the true length is $n + x$, where n is a whole number and x is between $\pm \frac{1}{2}$. The object is placed on the scale in an arbitrary position, and the positions of its ends are read. One end is at $m + y$, where m is an integer and y is between $\pm \frac{1}{2}$. Then the other end is at $m + n + x + y$. The position of the first end is then read in any case as m units. That of the other is read as $m + n - 1$ if $x + y$ is less than $-\frac{1}{2}$, as $m + n$ if $x + y$ is between $-\frac{1}{2}$ and $+\frac{1}{2}$, and as $m + n + 1$ if $x + y$ is greater than $+\frac{1}{2}$. x is fixed, but y is equally likely to have any value from $-\frac{1}{2}$ to $+\frac{1}{2}$. Thus the probability of a value of y between y_1 and y_2 is $y_2 - y_1$. We see that the measured length will be

$$\begin{aligned} &n - 1 \text{ if } y < -\tfrac{1}{2} - x, \\ &\quad n \quad \text{if } -\tfrac{1}{2} - x < y < \tfrac{1}{2} - x, \\ &n + 1 \text{ if } \tfrac{1}{2} - x < y. \end{aligned}$$

If x is positive the first alternative cannot arise, since y cannot be less than $-\frac{1}{2}$; for the second, the range of values

of y is from $-\frac{1}{2}$ to $\frac{1}{2}-x$, or $1-x$ in all; for the third, the range is x. Hence if x is positive the probability of an observed length equal to n units is $1-x$, and that of one equal to $n+1$ units is x. Similarly if x is negative the probability of an observed length equal to n units is $1+x$, and that of one of $n-1$ units is $-x$. In each case the possible measured lengths are the *two* multiples of the step adjacent to the true length.

5·3. Now consider the case where a large number of independent contributory causes affect the observed value. Suppose that the error ξ is given by

$$\xi = a_1\epsilon_1 + a_2\epsilon_2 + \ldots + a_n\epsilon_n, \qquad (1)$$

where $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ can all vary independently. Suppose that the probability that in any given trial ϵ_r will lie in a given range $d\epsilon_r$ is $E_r(\epsilon_r)\,d\epsilon_r$. Then the probability of a set of values within ranges $d\epsilon_1, d\epsilon_2, \ldots d\epsilon_n$ is

$$E_1(\epsilon_1)\,E_2(\epsilon_2)\ldots E_n(\epsilon_n)\,d\epsilon_1 d\epsilon_2 \ldots d\epsilon_n. \qquad (2)$$

We require the probability that ξ shall lie in a range ξ_1 to ξ_2. This is

$$I = \iiint \ldots \int E_1(\epsilon_1)\,E_2(\epsilon_2)\ldots E_n(\epsilon_n)\,d\epsilon_1 \ldots d\epsilon_n, \qquad (3)$$

where the range of integration is such that all values of each variable are permitted, subject to

$$\xi_1 \leqslant a_1\epsilon_1 + \ldots + a_n\epsilon_n \leqslant \xi_2. \qquad (4)$$

Now Heaviside's unit function $H(\xi)$, which is equal to 0 for ξ negative and 1 for ξ positive, is given by*

$$H(\xi) = \frac{1}{2\pi\iota}\int_{c-\iota\infty}^{c+\iota\infty}\frac{e^{\kappa\xi}}{\kappa}\,d\kappa. \qquad (5)$$

Also $\quad H(\xi-\xi_1) - H(\xi-\xi_2) = 1$ if $\xi_1 < \xi < \xi_2$, $\qquad (6)$

and otherwise $= 0$. Then

$$I = \frac{1}{2\pi\iota}\int_{c-\iota\infty}^{c+\iota\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\frac{e^{\kappa\xi}}{\kappa}\left(e^{-\kappa\xi_1} - e^{-\kappa\xi_2}\right)E_1(\epsilon_1)\,E_2(\epsilon_2)\ldots$$
$$\ldots E_n(\epsilon_n)\,d\kappa\,d\epsilon_1 \ldots d\epsilon_n, \qquad (7)$$

* Jeffreys, *Operational Methods in Mathematical Physics*, 1927.

where the ϵ's may now range over all real values independently. Now put

$$\int_{-\infty}^{\infty} e^{a_r \kappa \epsilon_r} E_r(\epsilon_r)\, d\epsilon_r = \Omega_r(a_r \kappa). \tag{8}$$

Then

$$I = \frac{1}{2\pi\iota} \int_{c-\iota\infty}^{c+\iota\infty} \frac{1}{\kappa} (e^{-\kappa\xi_1} - e^{-\kappa\xi_2})\, \Omega_1(a_1\kappa)\, \Omega_2(a_2\kappa) \ldots \ldots \Omega_n(a_n\kappa)\, d\kappa. \tag{9}$$

Now replace ξ_1 by ξ and ξ_2 by $\xi + d\xi$, and put

$$\Omega(\kappa) = \Omega_1(a_1\kappa)\, \Omega_2(a_2\kappa) \ldots \Omega_n(a_n\kappa). \tag{10}$$

Then I, the probability that ξ lies in a given range $d\xi$, becomes $P(\xi)\, d\xi$, where

$$P(\xi) = \frac{1}{2\pi\iota} \int_{c-\iota\infty}^{c+\iota\infty} e^{-\kappa\xi}\, \Omega(\kappa)\, d\kappa. \tag{11}$$

Now

$$\Omega_r(\theta) = \int_{-\infty}^{\infty} e^{\epsilon_r \theta} E_r(\epsilon_r)\, d\epsilon_r = \int_{-\infty}^{\infty} \left(1 + \theta\epsilon_r + \frac{\theta^2 \epsilon_r^2}{2!} + \ldots\right) E_r(\epsilon_r)\, d\epsilon_r$$

$$= 1 + \sum_{k=1}^{\infty} \frac{\theta^k}{k!} s_{rk}, \tag{12}$$

where

$$s_{rk} = \int_{-\infty}^{\infty} \epsilon_r{}^k\, E_r(\epsilon_r)\, d\epsilon_r. \tag{13}$$

Now form $\log \Omega_r(\theta)$, so that

$$\log \Omega_r(\theta) = \sum_{k=1}^{\infty} \frac{\theta^k}{k!} p_{rk}. \tag{14}$$

Then

$$\log \Omega(\kappa) = \sum_{r=1}^{n} \sum_{k=1}^{\infty} \frac{a_r{}^k \kappa^k}{k!} p_{rk} = \sum_{k=1}^{\infty} P_k \frac{\kappa^k}{k!}, \tag{15}$$

and

$$P(\xi) = \frac{1}{2\pi\iota} \int_{c-\iota\infty}^{c+\iota\infty} \exp\left(-\kappa\xi + \sum_{k=1}^{\infty} P_k \frac{\kappa^k}{k!}\right) d\kappa. \tag{16}$$

So far nothing has been assumed about the quantities s_{rk} except that $E_r(\epsilon_r)$ decreases for large absolute values of ϵ_r with sufficient rapidity to make the various integrals and series converge. We can, however, make all the s_{r1} zero by a

change of variable. For if this relation is not already satisfied we take a new ϵ_r' equal to $\epsilon_r - s_{r1}$, and then

$$\int_{-\infty}^{\infty} E_r(\epsilon_r)(\epsilon_r - s_{r1})\, d\epsilon_r = 0, \tag{17}$$

since

$$\int_{-\infty}^{\infty} E_r(\epsilon_r)\, d\epsilon_r = 1, \tag{18}$$

it being certain that ϵ_r lies between $\pm\infty$. If then we use ϵ_r' instead of ϵ_r the new s_{r1} is zero. Then p_{r1} and P_1 are 0. Also we define σ_r, the mean square or *standard* value of ϵ_r', by

$$\sigma_r^2 = \int_{-\infty}^{\infty} E_r(\epsilon_r)\, \epsilon_r'^2\, d\epsilon_r = s_{r2}. \tag{19}$$

We need no longer write accents, all component errors being supposed transformed in this way. We see that σ_r^2 is always positive; by convention we give σ_r the positive sign. Then

$$p_{r2} = \sigma_r^2; \quad P_2 = \sum_{r=1}^{n} a_r^2 \sigma_r^2; \tag{20}$$

$$P(\xi) = \frac{1}{2\pi\iota} \int_{c-\iota\infty}^{c+\iota\infty} \exp\left(-\kappa\xi + \tfrac{1}{2}P_2\kappa^2 + \sum_{\kappa=3}^{\infty} P_k \frac{\kappa^k}{k!}\right) d\kappa. \tag{21}$$

The integral is in a form suitable for evaluation by the method of steepest descents. If we omit the terms with $\kappa \geqslant 3$, there is a saddle point where

$$\kappa = \xi/P_2, \tag{22}$$

and the path of steepest descent is parallel to the imaginary axis, since P_2 is real and positive. Hence the integral reduces to

$$P(\xi) = (2\pi P_2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\xi^2/P_2). \tag{23}$$

Appreciable contributions to the integral arise only for values of $\kappa - \xi/P_2$ of order $(2/P_2)^{\frac{1}{2}}$ at most.

In most ordinary cases $E_r(\epsilon_r) = 0$ or is insignificant for values of ϵ_r much greater than σ_r. Then s_{rk} is of order σ_r^k; so is p_{rk}; and

$$P_k = \sum_{r=1}^{n} a_r^k p_{rk} = O\left\{\sum_{r=1}^{n} (a_r\sigma_r)^k\right\}, \tag{24}$$

and is of order $na^k\sigma^k$ if the $a_r\sigma_r$ are all comparable. Within the range where $\exp(-\kappa\xi + \frac{1}{2}P_2\kappa^2)$ is appreciable, then,

$$P_k\kappa^k = O(na^k\sigma^k)\left\{\frac{\xi}{P_2} \pm \left(\frac{2}{P_2}\right)^{\frac{1}{2}}\right\}^k, \qquad (25)$$

and if ξ is not large compared with $P_2^{\frac{1}{2}}$ this is

$$O(na^k\sigma^k)(P_2)^{-\frac{1}{2}k} = O(na^k\sigma^k)/(na^2\sigma^2)^{-\frac{1}{2}k} = O(n^{1-\frac{1}{2}k}). \quad (26)$$

If then n is large, $P_k\kappa^k$ is small throughout the neighbourhood of the saddle point for all values of $k > 2$; and then (23) is a close approximation to the true value of $P(\xi)$. Further, it makes

$$\int_{-\alpha}^{\alpha} P(\xi)\, d\xi = 1, \qquad (27)$$

nearly, where α is a moderate multiple of $(2P_2)^{\frac{1}{2}}$; and therefore values of ξ outside the range $\pm\alpha$ have an insignificant probability*.

We notice that the proof depends for its validity on the condition that when $k \geqslant 3$, P_k is small compared with $(P_2)^{\frac{1}{2}k}$, or

$$\sum_{r=1}^{n} a_r{}^k\sigma_r{}^k \text{ is small compared with } \left(\sum_{r=1}^{n} a_r{}^2\sigma_r{}^2\right)^{\frac{1}{2}k}. \quad (28)$$

If all the $a_r\sigma_r$ are equal or comparable, and n is large, this is true. But if one of them, for $r = m$ say, is so large as to contribute the greater portion of P_2, then P_2 is of order $a_m{}^2\sigma_m{}^2$, P_k is nearly $a_m{}^k\sigma_m{}^k$, and P_k is of the same order as $(P_2)^{\frac{1}{2}k}$. In such a case the normal law breaks down; it is indeed

* This discussion is taken mainly from Whittaker and Robinson's *Calculus of Observations*, which gives references to earlier writers. It has been modified by the introduction of Heaviside's unit function and the method of steepest descents. The consideration (27) is part of the proof. Our argument shows that in the specified conditions the terms in P_k for $k \geqslant 3$ do not matter on the path through the saddle point considered, by the usual considerations involved in the method. But if ξ was large compared with $(2P_2)^{\frac{1}{2}}$ our equation (26) would not hold. Then, however, we can use the fact that the total probability of all values of ξ is 1, and (27) shows that nearly all of it arises from such values as do make the approximation valid; when the probability is appreciable our result is correct, and when it is inappreciable our result is still correct.

obvious that then $a_m P(a_m \epsilon_m)$ is nearly $E_m(\epsilon_m)$. But when the contributions to P_2 arise in comparable amounts from a large number of the component errors the condition is true, and the normal law holds.

Suppose that a large number of the ϵ_r are of comparable importance, that is, that the probable range of variation of $a_r \epsilon_r$ is of the same order of magnitude for all of them, and that the others give smaller contributions to ξ. Then it is shown that the probability that ξ lies in a range $d\xi$ is $P(\xi)\, d\xi$, where

$$P(\xi) = \frac{h}{\sqrt{\pi}} e^{-h^2\xi^2}, \tag{29}$$

and h is a constant called the *modulus of precision*. This is called the *normal law of errors*. The probability of an error between o and ξ is

$$\int_0^{\xi} P(\xi)\, d\xi = \tfrac{1}{2} \operatorname{erf} h\xi. \tag{30}$$

When $h\xi$ is equal to 0·477, $\operatorname{erf} h\xi = \frac{1}{2}$. The corresponding value of ξ, equal to $0{\cdot}477/h$, is called the *probable error*; it has the property that the error is as likely to fall short of it as to exceed it. The *mean square* or *standard* error is defined by

$$\sigma^2 = \int_{-\infty}^{\infty} \xi^2 P(\xi)\, d\xi = \frac{1}{2h^2} = P_2 = \sum_{r=1}^{n} a_r^2 \sigma_r^2. \tag{31}$$

The probable error is 0·674 times the standard error.

5·31. There has been much discussion about the validity of the normal law of error. It on the whole follows the same lines as that associated with Laplace's theory of sampling: just as it is doubted whether there is any reason to believe Laplace's assumption that all compositions of the original class are equally probable, so it is doubted whether there is any reason to believe that errors are actually distributed according to the normal law. The solution in both cases seems to be much the same. If certain conditions are satisfied the normal law is definitely right; in other cases it is definitely

untrue. We have already had two simple cases where it is untrue. When an observation is made to the nearest multiple of the step of the instrument the error is the difference between the true value and that multiple, and is always the same. When a length or an interval of time is measured as the difference between two measures each made to the nearest multiple of the step, the possible observed values are the two nearest multiples of the step, and no others. In each case the normal law is simply inapplicable. But when the error arises as the resultant of a large number of independent errors of comparable importance the normal law is right. Two such cases are common.

5·32. Suppose that we make several observations of the same kind, of number n, and that we take the mean. Then each observation is liable to an error of the same magnitude, and the standard value of each is the same. The mean is $1/n$ times the sum of the individual errors, so that each a_r in the foregoing discussion is $1/n$, and

$$\sigma^2 = \Sigma \left(\frac{\sigma_r}{n}\right)^2 = \frac{1}{n^2} \Sigma \sigma_r^2. \qquad (1)$$

The conditions for the validity of the normal law hold if n is large. If for instance we consider the measure of a length, when the step is unity, and the true value is an integer $+ x_r$, where x_r is positive, $E_r(\epsilon_r) = 0$ unless ϵ_r is either $-x_r$ or $1 - x_r$. The probability that ϵ_r is $-x_r$ is $1 - x_r$; the probability that ϵ_r is $1 - x_r$ is x_r. Then $\int_a^b E_r(\epsilon_r)\, d\epsilon_r = 0$ unless the range a to b includes either $-x_r$ or $1 - x_r$; if it does include $-x_r$ the integral is $1 - x_r$; if it includes $1 - x_r$ the integral is x_r, however short the range may be*.

Then

$$\sigma_r^2 = \int_{-\infty}^{\infty} E_r(\epsilon_r)\, \epsilon_r^2\, d\epsilon_r$$
$$= (1 - x_r)\, x_r^2 + x_r (1 - x_r)^2 = x_r (1 - x_r). \qquad (2)$$

* Stieltjes integrals are understood. Cf. Hobson, *Theory of Functions of a Real Variable*, 1, 507.

When n is large and the x_r are regularly distributed from 0 to 1,

$$\Sigma \sigma_r^2 = \Sigma x_r (1 - x_r)$$
$$= n \int_0^1 x (1 - x)\, dx, \text{ nearly,}$$
$$= \tfrac{1}{6} n, \qquad (3)$$

provided n is large enough for the theory of sampling to be applicable. Then

$$\sigma^2 = \frac{1}{6n}; \quad P(\xi) = \frac{h}{\sqrt{\pi}} e^{-h^2\xi^2} = \sqrt{\frac{3n}{\pi}}\, e^{-3n\xi^2}. \qquad (4)$$

Strictly speaking the possible values of the mean are all multiples of the step divided by n, but this gives no trouble provided that we consider only the probabilities of errors within ranges greater than $1/n$.

This theory is not applicable if the *same* length is measured several times, for then σ_r is always the same and a function of x, ranging, by (2), from 0 for $x = 0$ or 1 to $\frac{1}{2}$ for $x = \frac{1}{2}$. The condition that the errors must be independent is then not satisfied. We notice that in this case

$$\int_{-\infty}^{\infty} E(\epsilon)\, \epsilon\, d\epsilon = -(1 - x)\, x + x(1 - x) = 0. \qquad (5)$$

5·33. Another case where the normal law appears to hold is one where considerable attention has been given to possible sources of error and all the most serious ones have been traced, as in many astronomical observations. The remaining ones then probably contain several just below the limit of what can be detected individually, and the normal law will hold approximately.

5·34. The normal law is true unless there are one or a few sources of error of sufficient importance to dominate all the rest. But if there is a single main source of error we should

still consider its probable distribution. It may be one of the types already considered, arising from the step of the instrument. If so its properties may be considered known. It may be the result of a definite mistake on the part of the observer, as when an astronomer observing a meridian transit makes a miscount of a second. Criteria for detecting such mistakes are needed; at present we notice only that they are capable of giving errors of certain discrete values, which are multiples of the step. Other factors not allowed for may have similar properties; that is, they affect only a small fraction of the observations, but when they do arise they give errors larger than are usual. Such errors may be capable of only one sign; thus the astronomer may occasionally count too few seconds, but never too many.

There may on the other hand be a single source of error capable of giving many different values. There may for instance be an unknown periodic disturbance. The practical solution here is that the periodic character of the residuals is noticed, and its amount can be determined by harmonic analysis and allowed for; its cause then becomes a matter for independent inquiry. Such a case arose in the discovery by Chandler of the 14-monthly and annual terms in the variation of latitude. But such individual sources of error may have many different distributions of probability; and in practice the issue is very like that of assessing the distribution of prior probability in the theory of sampling. We start from a state of ignorance such that all observed values of the variable are equally probable. By experience we build up knowledge that the observed values are concentrated in a short range about the value given by a simple law, and by studying all our previous knowledge about modes of distribution of errors we could, given sufficient trouble, assess the probabilities of given errors. But the effort would be more trouble than it is worth. In practice it is better to take a sufficiently large number of observations to make the posterior probability practically independent of the prior probability.

5·4. In practice we are not much interested in the errors as such, except in so far as they may show a systematic character that may repay special investigation. What we want is the true value, and if we cannot find it, we want to choose an adopted value as near as possible to it. That is, given the observed values, we wish to assess the probabilities that the true values may lie in various ranges. The problem therefore becomes one of inverse probability, and the prior probabilities of different true values must be taken into account.

5·41. Consider first the case of a single reading made to the nearest multiple of the step; the observed value is n, where the step is the unit. The true value is $n + x$. Then the prior probability that x may lie within a range is proportional to the length of the range; if $P(x)\,dx$ is the prior probability that x lies within a range dx, $P(x)$ is a constant. The probability of getting the reading n is 1 when x is between $\pm \frac{1}{2}$ and zero when x is outside that range, for then another integral value would be read. Hence the posterior probability that x lies within a range dx is

$$\frac{P(x)\,dx \,.\, 1}{\int_{-\infty}^{\infty} P(x)\,dx} = dx \text{ when } -\tfrac{1}{2} < x < \tfrac{1}{2},$$

$$\frac{P(x)\,dx \,.\, 0}{\int_{-\infty}^{\infty} P(x)\,dx} = 0 \text{ when } x < -\tfrac{1}{2}.$$

Thus after the observation, or any number of such observations, the posterior probability of x is uniformly distributed between $\pm \frac{1}{2}$.

5·42. Consider next a length or a time interval determined by difference. The observed values are l equal to n and m equal to $n + 1$. The true value is $n + x$, and $P(x)$ is constant. For a given x the probability of a reading n is $1 - |x|$ when x is between ± 1 and otherwise zero; the probability of a

reading $n+1$ is $1-|1-x|$ when $1-x$ is between ± 1 and otherwise zero. If then x was negative the readings $n+1$ would not arise; if x was greater than 1 the readings n would not arise. For $0<x<1$, the probability of l readings equal to n and m equal to $n+1$ is ${}^{l+m}C_l\,(1-x)^l\,x^m$. The posterior probability that x lies in a range dx is therefore 0 for $x<0$ or $x>1$, and when $0<x<1$ is

$$\frac{P(x)\,dx\,{}^{l+m}C_l\,(1-x)^l\,x^m}{\int_0^1 P(x)\,dx\,{}^{l+m}C_l\,(1-x)^l\,x^m} = \frac{(1-x)^l\,x^m\,dx}{\int_0^1 (1-x)^l\,x^m\,dx}$$

$$= \frac{(1-x)^l\,x^m}{B(l+1,m+1)}\,dx = \frac{(l+m+1)!}{l!\,m!}(1-x)^l\,x^m\,dx. \quad (1)$$

The coefficient of dx is a maximum when

$$x = \frac{m}{l+m}, \quad (2)$$

so that the mean of the observations is the most probable value. Calling this value x_0, we find easily that when l and m are large the posterior probability is proportional to

$$\exp\left\{-\frac{(l+m)^3}{2lm}(x-x_0)^2\right\}dx.$$

Thus if we take x_0 as the adopted value, the probabilities of different true values are distributed about x_0 according to the normal law, with a standard deviation σ_a given by

$$\sigma_a{}^2 = \frac{lm}{(l+m)^3}. \quad (3)$$

We notice the advantage of this method over direct reading. When a single quantity has to be measured as the nearest multiple of the step, the same observation may be made an indefinite number of times without in the least affecting the precision of the adopted value. But when it is determined by difference and the measure is repeated a large number of times, the standard difference between the adopted and true values may be reduced indefinitely.

5·43. When the normal law of error applies, we proceed as follows. The true value being now taken as x, and the observed value as $x+\xi$, then the probability of an observed value in a range $d\xi$ is $\frac{h}{\sqrt{\pi}} e^{-h^2\xi^2}\, d\xi$. The probability of a set of errors in ranges about $\xi_1, \xi_2, \ldots \xi_n$ is then

$$\left(\frac{h}{\sqrt{\pi}}\right)^n \exp\{-h^2(\xi_1^2+\xi_2^2+\ldots+\xi_n^2)\}\, d\xi_1 d\xi_2 \ldots d\xi_n. \quad (1)$$

But actually both x and h are initially unknown, and we are trying to find x from the observed values. Calling these $x_1, x_2, \ldots x_n$, we have

$$\xi_1 = x_1 - x, \ldots \xi_n = x_n - x, \quad (2)$$

$$d\xi_1 = dx_1, \ldots\ldots d\xi_n = dx_n. \quad (3)$$

If the prior probability that x and h lie simultaneously in ranges dx, dh is $P(x,h)\,dx\,dh$, the posterior probability that they lie in these ranges is

$$\frac{P(x,h)h^n \exp[-h^2\{(x_1-x)^2+\ldots+(x_n-x)^2\}]\,dx\,dh}{\int_{-\infty}^{\infty}\int_0^{\infty} P(x,h)h^n \exp[-h^2\{(x_1-x)^2+\ldots+(x_n-x)^2\}]\,dx\,dh}, \quad (4)$$

the factor $\pi^{-\frac{1}{2}n}\, dx_1 dx_2 \ldots dx_n$ being the same for all values of x and h.

As usual the posterior probability depends on the prior probability. In most cases the prior probability of x is nearly uniformly distributed, at any rate over a range several times that covered by the observations. We are initially prepared for values of x over a wide range, and the purpose of making observations at all is to permit a considerable reduction of this range. The position is different with regard to h. Initially we may have no special views about the probability of one value of h rather than another, but we do at least know that negative values are excluded, since they would imply negative probabilities. Again, x is not usually in fact a number; it is usually a length or an interval of time, and h is a reciprocal

of whatever kind of magnitude x is, while the standard error σ is the same kind of quantity as x. There seems to be no special reason for measuring the precision in terms of h rather than σ, and their product is constant, so that

$$d \log h + d \log \sigma = 0. \qquad (5)$$

If then $P(x, h)\, dh$ is proportional to dh/h or $d\sigma/\sigma$, an ambiguity is removed. It means that the probability of a value of σ or h within a definite range is proportional to the increase of its logarithm; if $h_1/h_2 = h_3/h_4$, h is as likely to lie between h_1 and h_2 as between h_3 and h_4. The probability of a value of h within any range is then independent of any scale of measurement; it is distributed in the same way among different values whatever our units. If any other function of h was chosen we should be assigning a definite prior probability to a value of h less than a certain quantity, and this would put a particular value of a physical quantity in a privileged position *a priori*. In many cases, then, it seems reasonable to take $P(x, h)$ proportional to $1/h$.

This is not, however, quite a complete statement, because it makes $\int_0^\infty P(x, h)\, dh$ diverge at both limits. To make this integral equal to 1 we should therefore have to include a zero factor unless very small and very large values of h are excluded. This does appear to be the case. We choose the length of our scale so that all the measures will be included within it easily; that is, all the important values of h are large compared with the reciprocal of the length of the scale. Again, if the scatter of the observations is comparable with the step of the scale, the finiteness of the step is a dominant source of error and the normal law does not apply at all. We are therefore restricted to a range of values of h that make σ large compared with the step of the scale and small compared with the length of the scale. The range of admissible values of $\log h$ is now large but finite, and within this range we may suppose their prior probabilities distributed uniformly except near the ends.

We now introduce the mean value, defined by

$$nx_0 = x_1 + x_2 + \ldots + x_n, \tag{6}$$

and write

$$x_1 - x = (x_1 - x_0) + (x_0 - x), \tag{7}$$

and so on; then

$$(x_1 - x)^2 + (x_2 - x)^2 + \ldots + (x_n - x)^2 = (x_1 - x_0)^2 + (x_2 - x_0)^2 + \ldots + (x_n - x_0)^2 + n(x - x_0)^2. \tag{8}$$

The quantities $x_1 - x_0$ and so on are the residuals, ξ_1' say, and $x_0 - x$ is the error of the mean value. Then

$$\int_{-\infty}^{\infty} \exp\{-nh^2(x - x_0)^2\}\, dx = \frac{1}{h}\left(\frac{\pi}{n}\right)^{\frac{1}{2}}, \tag{9}$$

and if we denote the posterior probability of values of x and h in the range $dx\,dh$ by $I(x, h)\,dx\,dh$ we have

$$I(x, h) = \left(\frac{n}{\pi}\right)^{\frac{1}{2}} \frac{h^{n-1} \exp\left[-h^2\{\Sigma \xi'^2 + n(x - x_0)^2\}\right]}{\int_0^{\infty} h^{n-2} \exp\left[-h^2 \Sigma \xi'^2\right] dh}. \tag{10}$$

Put

$$\Sigma \xi'^2 = n\sigma'^2 \tag{11}$$

so that σ' is the standard residual. We have

$$\int_0^{\infty} h^{n-2} \exp(-\alpha^2 h^2)\, dh = \frac{1}{2\alpha^{n-1}} \Pi\{\tfrac{1}{2}(n - 3)\}, \tag{12}$$

and

$$I(x, h) = 2\left(\frac{n}{\pi}\right)^{\frac{1}{2}} \frac{(n\sigma'^2)^{\frac{1}{2}(n-1)}}{\Pi\{\frac{1}{2}(n-3)\}} h^{n-1} \exp\left[-nh^2\{\sigma'^2 + (x - x_0)^2\}\right], \tag{13}$$

the large values of h making an inappreciable contribution in any case, and the small ones if $n > 1$. $I(x, h)$ does not break up into two factors, one a function of x and the other of h, so that it would not be correct to speak, on the data, of the probabilities of given values of $x - x_0$ and h separately.

The probability of a value of x in the range dx, irrespective of h, is

$$dx \int_0^\infty I(x, h)\, dh$$
$$= 2 \left(\frac{n}{\pi}\right)^{\frac{1}{2}} \frac{(n\sigma'^2)^{\frac{1}{2}(n-1)}}{\Pi\{\frac{1}{2}(n-3)\}} \frac{\Pi\{\frac{1}{2}(n-2)\}}{2\,[n\{\sigma'^2 + (x-x_0)^2\}]^{\frac{1}{2}n}}\, dx$$
$$= \frac{1}{\sqrt{\pi}} \frac{\Pi\{\frac{1}{2}(n-2)\}}{\Pi\{\frac{1}{2}(n-3)\}} \frac{\sigma'^{n-1}}{\{\sigma'^2 + (x-x_0)^2\}^{\frac{1}{2}n}}\, dx, \qquad (14)$$

so that the posterior probability of x is not distributed according to the normal law.

But if n is large, and $x - x_0$ small compared with σ',

$$\{\sigma'^2 + (x-x_0)^2\}^{\frac{1}{2}n} = \sigma'^n \exp \tfrac{1}{2}n \left(\frac{x-x_0}{\sigma'}\right)^2, \qquad (15)$$

nearly, and the probability of a given value of x is proportional to $\exp\{-n(x-x_0)^2/2\sigma'^2\}$. The mean value is in any case the most probable; in this case the probabilities of the true values are distributed about it according to the normal law with a standard deviation $\sigma'/\sqrt{n}$. Subject to the same condition we can put $(x-x_0)^2$ equal to its standard value σ'^2/n in $I(x, h)$; then

$$I(x, h) \propto h^{n-1} \exp\left(-nh^2\sigma'^2 \frac{n+1}{n}\right)$$
$$= h^{n-1} \exp\{-(n+1)h^2\sigma'^2\}. \qquad (16)$$

This is now independent of x, and may be taken to give the distribution of probability of h. It is a maximum when

$$h^2\sigma'^2 = \frac{1}{2}\frac{n-1}{n+1}, \qquad (17)$$

so that the most probable value of h is nearly $1/\sqrt{2}\sigma'$. Near this value of h, h_0 say, the probabilities are distributed nearly according to the law

$$I(x, h) \propto \exp\{-2(n+1)\sigma^2(h-h_0)^2\}. \qquad (18)$$

The standard deviation of h is $(n+1)^{-\frac{1}{2}}/2\sigma'$.

But we must remember that it is only in a rough sense that we can speak of the posterior probabilities of values of x and h separately even when n is large. If we put $(x - x_0)^2$ equal to 0, corresponding to the most probable value of $x - x_0$, instead of to its standard value, the resulting probabilities of h would be somewhat differently distributed. The most probable value of h and its standard deviation are strictly functions of $x - x_0$.

5·5. The most commonly quoted proof of the normal law of error is that of Gauss, which appears to show that if the mean is the most probable value the errors must follow the normal law. A case has arisen above where the mean is the most probable value and the errors do not follow the normal law. It is therefore desirable to reconsider Gauss's argument and see where the difference has entered. He proceeds by assuming that the true value is x, and that the probability of an observation within a range dx_1 about x_1 is $\phi(x_1 - x)\,dx_1$. Then the probability of a set in the ranges $dx_1, dx_2, \ldots dx_n$ is

$$\phi(x_1 - x)\,\phi(x_2 - x) \ldots \phi(x_n - x)\,dx_1 dx_2 \ldots dx_n. \quad (1)$$

Given the observed values, then, the probability of a value of x is proportional to

$$\phi(x_1 - x)\,\phi(x_2 - x) \ldots \phi(x_n - x)\,dx, \quad (2)$$

if the prior probability of x is uniformly distributed. This is a maximum for variations in x if

$$\frac{d}{dx}\log\phi(x_1 - x) + \frac{d}{dx}\log\phi(x_2 - x) + \ldots + \frac{d}{dx}\log\phi(x_n - x) = 0. \quad (3)$$

But the postulate that the mean value is the most probable says that this condition must be equivalent to

$$(x_1 - x) + (x_2 - x) + \ldots + (x_n - x) = 0, \quad (4)$$

for all values of the differences, and therefore

$$\frac{\frac{d}{dx}\log\phi(x_1-x)}{x_1-x}=\frac{\frac{d}{dx}\log\phi(x_2-x)}{x_2-x}=\ldots=\frac{\frac{d}{dx}\log\phi(x_n-x)}{x_n-x} \quad (5)$$

$$= 2h^2, \quad (6)$$

say, since each ratio is the same and therefore cannot vary with x. Integrating we find that

$$\phi(x_1-x) \propto \exp\{-h^2(x_1-x)\}^2, \quad (7)$$

which is the normal law of error.

This mistake is in the equation (1), which supposes that the probability of getting all the observations $x_1, x_2, \ldots x_n$ is the product of the probabilities of each observation separately. It supposes, that is, that when the observations $x_1, x_2, \ldots x_{n-1}$ have been made the probability that x_n will have a certain value is just what it was at the start. It therefore constitutes another contradiction of the principle that it is possible to learn from experience. If the early observations are found to have a small scatter, the next will be expected to be near them; if they have a large scatter we shall correspondingly expect the next to deviate considerably from the mean of those already made. If they all repeat one of two constant values, we shall expect the next to have one of those values. Gauss's proof is in fact valid if we know beforehand all about the distribution of the probability of error; it is inapplicable when it is from the observations themselves that we are trying to find this distribution.

When we say that "the normal law holds" we mean that there are true values of x and h such that the errors satisfy the normal law. In the usual practical case the possible values of x and h are scattered over a wide range, the normal law holding for each pair of values. If we try to assess the total prior probability of an observed value x_1 for a given x, by adding up the contributions for all values of h, the

result is not of the normal form; if $P(x, h)$ is proportional to $1/h$, the prior probability of a given ξ is proportional to

$$\frac{d\xi}{\sqrt{\pi}} \int_0^\infty e^{-h^2\xi^2}\, dh,$$

which is not proportional to $e^{-h_0^2\xi^2}$ for any value of h_0. Similarly the posterior probabilities are not of the normal form, even when the normal law holds. It is the *component* probability from each pair of values of x and h that is referred to when we speak of the normal law of error; any attempt to compound probabilities destroys the normal form.

5·6. In addition to errors with probabilities following the normal law and those arising from the step of the instrument many other types exist. The probability of a given distribution of error, before the observations are taken, is in each case quite definite, but involves taking into account the whole of our previous knowledge about what distributions of error have occurred in the past. Its calculation would be overwhelmingly laborious, and the effect on the result would in most practical cases not be worth while. If there is no strong and obvious reason to expect any particular law of error in a given case, there is no better plan than to take a large number of observations and draw a smooth curve to represent the frequency of their departures from some convenient standard value. But the question arises, what in this case is the most probable value? The answer will depend on the circumstances. If the observations show a strong tendency to collect about two definite values, that fact is evidence that the errors arise from some disturbing factor with a finite step, and we cannot do better than to take the arithmetic mean. They may be approximately symmetrically distributed about the mean value; in that case also, if there is no previous reason to expect the errors to be predominantly of one sign, we may take the mean value as the most probable. But it may turn out that their distribution is noticeably asymmetrical. The observations on one side of the mean may be few, but with

large deviations, while those on the other side are many with small deviations. In that case the placing of the most probable value with respect to the mean requires either assessment of the prior probability or special examination of the actual causes of the errors. If neither is carried out an uncertainty about the position of the most probable value necessarily remains. Three alternatives are usually considered in such a case: the arithmetic mean, the mode, and the median. The median is defined by the condition that as many observations exceed it as fall short of it; the mode is such that the number of observed values for a given range is greatest there. In general with asymmetrical distributions all three are different. The median would be the most probable if a positive error is as likely as a negative one, irrespective of their magnitudes; the arithmetic mean may be the most probable if the magnitudes of the errors matter. Both alternatives may arise in different cases. The mode is the most probable if there is some reason to expect that the errors arise from special causes not present at all in the majority of the observations. The use of the mean as the adopted value has the advantage that it makes the standard deviation a minimum. The median has the advantage that we can divide the observed values, in order of magnitude, into four classes, each containing as nearly as possible the same number of observations. The median comes at the boundary between the two middle classes, while the extremes of the two middle classes specify a range such that a given observation is as likely to lie within it as outside it. In this sense such a classification determines a probable error; or rather two probable errors, one for positive and the other for negative errors. In no case will the probable error of a single observation, that of the adopted value, nor the mean square error, follow the same quantitative rules as have been determined for cases where the normal law holds.

If the only purpose of the observations is to determine a single quantity as accurately as possible, and the errors turn

out to be asymmetrically distributed, there seems to be nothing to do but to consider which of the conditions for the arithmetic mean, the median, and the mode is the most likely to be applicable in the given case, and to choose the adopted value accordingly. A method often considered in such a case is to attempt to allow for the terms in $P_3, P_4 \ldots$ and so on in 5·3 (21). Thus

$$\left(1 + \sum_{k=3}^{\infty} (-1)^k \frac{P_k}{k!} \frac{\partial^k}{\partial \xi^k}\right) \frac{1}{2\pi\iota} \int_{c-\iota\infty}^{c+\iota\infty} \exp\left(-\kappa\xi + \tfrac{1}{2} P_2 \kappa^2\right) d\kappa$$

$$= \frac{1}{2\pi\iota} \int_{c-\iota\infty}^{c+\iota\infty} \exp\left(-\kappa\xi + \tfrac{1}{2} P_2 \kappa^2 + \sum_{k=3}^{\infty} P_k \frac{\kappa^k}{k!}\right) d\kappa$$

$$= P(\xi), \text{ subject merely to a convergency condition,}$$

so that we can write

$$P(\xi) = \left(1 + \sum_{k=3}^{\infty} (-1)^k \frac{P_k}{k!} \frac{\partial^k}{\partial \xi^k}\right) \frac{1}{(2\pi P_2)^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2} \frac{\xi^2}{P_2}\right).$$

The expansion can then be carried out; the terms are known functions of ξ with adjustable coefficients involving the P_k. By an extension of the method used for finding the posterior probabilities of values of x and h when the normal law holds, we can now use the distribution of the observations to find both P_2 and the higher P_k as closely as possible, and still to estimate the distribution of the posterior probability among various values of x. But it seems to me that such a procedure can lead nowhere. The normal law is valid as it stands when there are a large number of component errors of comparable magnitude. If it requires modification, it is because there are one or a few sources of error of predominating importance, and the law of error is determined mainly by these. If the extended law is applied it will only lead back to the law of the dominant component error, whatever that may be; and if the observations cannot determine that directly no modification of the normal law will do so, for the condition for a few terms of the series to give an approximation to the whole

is not satisfied, all the terms in fact being of the same order of magnitude.

5·7. Another warning is needed with regard to the quantities s_{r1}. If these do not vanish, the error that we have shown in certain cases to follow the normal law does not arise directly from the actual component errors ϵ_r, but from their differences from their associated s_{r1}. In fact ξ' is not $\Sigma\, a_r\epsilon_r$, but

$$\xi' = \Sigma\, a_r\epsilon_r' = \Sigma\, a_r\epsilon_r - \Sigma\, a_r s_{r1} = \xi - \Sigma\, a_r s_{r1}.$$

The observed value is $x + \xi$, that is, $x + \sum_{r=1}^{n} a_r s_{r1} + \xi'$, where ξ' follows the normal law. Then however many observations we may use to determine our mean, the quantity that has the mean for its most probable value is not x, the true value, but $x + \sum_{r=1}^{n} a_r s_{r1}$. We can never find the true value from this without some knowledge of the sum $\sum_{r=1}^{n} a_r s_{r1}$, which affects every observation equally. It is usual to call such a constant error the *systematic* error, while the deviations ξ' that do satisfy the normal law are called *accidental* errors.

If now the probability of error is asymmetrically distributed, that means that errors of one sign are likely to be more frequent or larger than those with the other sign; in either case a systematic error is to be expected. In any case, that is, where the distribution is asymmetrical, the existence of a systematic error may be inferred. But it may exist also where the distribution is symmetrical. In either case the observations give us no means of evaluating it; this can be done only by way of other considerations.

At first glance the problem of systematic error seems to stultify our whole procedure; for it means that, however many observations we may take, the difference between the adopted value and the true value remains unknown. Yet we still have our assurance that the observed values nearly satisfied the physical law under test; the adopted value cannot be

far wrong. At the worst we could make it a convention to take the mean as the adopted value in the case of a quantity known to be nearly constant; or we could always find the parameters in a law by the method of least squares. The true value in any case does not differ much from the adopted value; the question at issue is *how* much it is likely to differ. This reduces to the task of evaluating the systematic error, which is in any case small, of the order, for instance, of the difference between the mean and the median. We may attempt to do this from previous knowledge, by measuring other variables directly and allowing for them; or we may determine the quantity under consideration by means of other laws that involve it and may give different systematic errors, and then compare the results. The differences may indicate the nature and extent of the systematic errors and suggest means of tracing them to their causes. In fact systematic errors are, and always will be, the curse of the present and the hope of the future.

5·71. We may be interested in the arithmetic mean for other reasons; for instance, it is wanted directly in the evaluation of an integral. Suppose that the true value is x, and the error of an observation ξ. Let the probability of an error between ξ and $\xi + d\xi$ be $E(\xi)\,d\xi$. We take

$$\int_{-\infty}^{\infty} E(\xi)\,d\xi = 1; \quad \int_{-\infty}^{\infty} E(\xi)\,\xi\,d\xi = s;$$

$$\int_{-\infty}^{\infty} E(\xi)\,(\xi - s)^2\,d\xi = \sigma^2, \quad (1)$$

so that

$$\int_{-\infty}^{\infty} E(\xi)\,\xi^2 d\xi = \sigma^2 + s^2. \quad (2)$$

Now consider a set of observations ξ_r, n in number, and their mean

$$\xi_0 = \frac{1}{n}\Sigma\,\xi_r. \quad (3)$$

Then the probability that the mean would be in a range $d\xi_0$ is

$$\frac{h}{\sqrt{\pi}} \exp\{-h^2(\xi_0 - s)^2\}\,d\xi_0, \quad (4)$$

provided the conditions of 5·3 are applicable. The standard error is the same for each observation, since all are made in the same conditions, and they are independent. Hence, if σ_0 is the standard error of the mean,

$$\sigma_0^2 = \sum_{r=1}^{n} \frac{1}{n^2} \cdot \sigma^2 = \frac{\sigma^2}{n}; \quad h^2 = 1/2\sigma_0^2. \tag{5}$$

The conditions required hold provided that n is large. The probabilities of the mean value are therefore distributed about $x + s$ according to the normal law even if those of the original observations are not. Further,

$$\sum_{r=1}^{n} (\xi - s)^2 = \sum_{r=1}^{n} \{(\xi - \xi_0)^2 + 2(\xi - \xi_0)(\xi_0 - s) + (\xi_0 - s)^2\}. \tag{6}$$

Of the terms on the right, the second is zero. The first can be found from the observations and denoted by $n\sigma'^2$, where σ' is the standard deviation. The last may be zero, but we may suppose, seeing that the standard value of $(\xi - s)^2$ is σ^2 and that of $(\xi_0 - s)^2$ is σ^2/n, that the ratio of the corresponding sums is $n : 1$. This is an approximation, which will sometimes exceed and sometimes fall short of the truth. Then we can take

$$n\sigma^2 \left(1 - \frac{1}{n}\right) = n\sigma'^2, \tag{7}$$

or

$$\sigma^2 = \frac{n}{n-1}\sigma'^2; \quad \sigma_0^2 = \frac{\sigma'^2}{n-1}. \tag{8}$$

This approximation is subject to the same sort of uncertainty as arose in dealing with the determination of the standard error when the normal law is satisfied. We have obtained in this way an estimate of the standard error of the arithmetic mean when the actual law of error is not the normal one.

5·72. It often happens that the quantity sought can be found from several different types of data. The mean distance of the sun, for instance, may be found from observations of

the transit of Venus, observations of Mars or an asteroid near opposition, from the moon's parallactic inequality, or from the aberration of light. Suppose that the true value is x, and that we have several methods of measurement. In the rth method the probability, given x, of an error between ξ_r and $\xi_r + d\xi_r$ is $E(\xi_r)\, d\xi_r$, for a single observation. Denote an individual observation in the rth series by ξ_{rs}, and suppose that there are n_r such observations. Consider the sum

$$\xi = \sum_r \sum_s a_r \xi_{rs} = \sum_r n_r a_r s_r + \sum_r \sum_s a_r (\xi_{rs} - s_r)$$
$$= \xi_0 + \xi', \qquad (1)$$

say, where the a_r are constants. Then in certain conditions the probability that ξ will lie in a given range $d\xi$ is $P(\xi)\, d\xi$, where

$$P(\xi) = \frac{h}{\sqrt{\pi}} e^{-h^2\xi'^2}, \qquad (2)$$

and
$$2h^2\sigma^2 = 1; \quad \sigma^2 = \sum_r \sum_s a_r^2 \sigma_r^2 = \sum_r n_r a_r^2 \sigma_r^2. \qquad (3)$$

When all the observations are equal we want ξ to have the same value. Hence we take

$$\sum_r \sum_s a_r = \sum_r n_r a_r = 1. \qquad (4)$$

Otherwise the a_r are at our disposal.

Suppose that we want to make σ^2 as small as possible. We introduce a multiplier λ and say that

$$\tfrac{1}{2} d\sigma^2 = \sum_r n_r (\sigma_r^2 a_r - \lambda)\, da_r = 0, \qquad (5)$$

for all da_r, if λ is chosen suitably. Hence

$$a_r = \frac{\lambda}{\sigma_r^2}; \quad \sum_r \left(\frac{n_r}{\sigma_r^2}\right) \lambda = 1, \qquad (6)$$

and finally
$$a_r = \sigma_r^{-2} \Big/ \sum_r \frac{n_r}{\sigma_r^2}, \qquad (7)$$

$$\sigma^2 = \sum_r n_r \lambda^2 / \sigma_r^2 = \lambda = 1 \Big/ \sum_r \frac{n_r}{\sigma_r^2}. \qquad (8)$$

But the σ_r are the standard errors of $x - s_r$ as found from the separate methods; if σ_r' is the observed standard deviation in each and σ_{r_0} the standard error of the mean value we have nearly

$$\frac{\sigma_r^2}{n_r} = \frac{\sigma_r'^2}{n-1} = \sigma_{r_0}^2. \tag{9}$$

If then we determine the standard error of each mean value as in 5·71 we have

$$\frac{1}{\sigma^2} = \sum_r \frac{1}{\sigma_{r_0}^2}. \tag{10}$$

The conditions for the validity of the normal law of error are that the individual errors shall be independent, which they are; and that the largest contributions to σ^2 from the individual errors shall arise in comparable amounts from a large number of them and not from only a few. The latter condition is satisfied by the terms in σ^2 arising from a single series of observations, and *a fortiori* from those from all the series together. The probabilities of errors in weighted means derived from several series of observations therefore satisfy the normal law; and the standard error can be computed from the means and standard errors of the separate series by the same methods as are applicable if the probabilities of error in each series are distributed according to the normal law.

The practice of weighting the means from the separate series according to the inverse squares of their standard errors is open to some objection, because it neglects the question of systematic error. The quantity that follows the normal law is not the actual error of the final mean, but this error less by

$$\sum_r n_r a_r s_r = \frac{\sum_r n_r s_r/\sigma_r^2}{\sum_r n_r/\sigma_r^2}. \tag{11}$$

To make the error of the final mean small we want not only to make its standard accidental error as small as possible, but also to reduce as far as we can its systematic error. To choose the a_r as in (7) achieves the first object; but there is little

ground for supposing that the same choice is suitable for the second object. In particular the a_r, as we have chosen them, depend on the number of observations in the series; the systematic errors do not. To put the matter in another way, the means derived from the separate series in general differ. The differences arise partly from the fact that the different methods give different systematic errors, and partly from accidental errors. The latter can be reduced indefinitely by taking enough observations, but no number of observations will reduce the systematic errors. If the means differ by amounts large compared with their standard errors, it is fair to infer that the differences arise from systematic error, and the weights assigned are illusory. If we have previous reason to expect systematic error from any method, its amount may be inferred from the differences between the mean given by that method and those given by the others. If all the methods are initially equally likely to have systematic errors of a given amount, we should take a simple unweighted mean, at any rate until the causes of the outstanding discrepancies have been investigated.

5·8. There is a common type of error, which arises from the co-operation of a large number of causes of comparable importance, together with one or a few that affect only a small fraction of the observations, but produce large errors when they do occur. One such example has been mentioned already, when an astronomer observing a transit makes a miscount of a second. Such a cause implies an incompleteness in the normal law of error and therefore casts doubt on the adoption of the arithmetic mean as the most probable value. We require a criterion for recognizing observations so affected when they occur. An absolute criterion is impossible, for a deviation of any magnitude is theoretically possible, even when the normal law applies and the standard error is known already. But errors greater than a moderate multiple of the standard error are so rare that we may say that when they arise they probably

come from some unusual cause. If so, we shall be justified in rejecting them and determining the adopted value and the standard error of the adopted value from the others. This course will sometimes be mistaken, because they may really arise as the large errors to be expected occasionally from the normal law itself; and if the normal law is applicable to the whole of the observations the most probable value is the arithmetic mean, and the mean after rejecting an observation is not the most probable value.

In the circumstances we are considering the error ξ is of the form $\xi_1 + \xi_2$, where ξ_1 follows the normal law. The probability of value of ξ_2 follows the law

$$E_2(\xi_2)\,d\xi_2 = \frac{m-1}{m} \text{ when } \xi_2 = 0 \text{ lies within } d\xi_2, \quad (1)$$

$$\int_{-\infty}^{\infty} E_2(\xi_2)\,d\xi_2 = \frac{1}{m} \text{ when a range about } \xi_2 = 0 \text{ is excluded}, \quad (2)$$

$$m\int_{-\infty}^{\infty} E_2(\xi_2)\,\xi_2^2\,d\xi_2 = \sigma_2^2 \text{ with the same restriction.} \quad (3)$$

Then the error ξ_2 arises in only $1/m$ of the cases, but if its standard value σ_2 is found from the cases when it can arise it much exceeds σ_1, the standard error of those observations that do follow the normal law.

In practice m and the form of $E_2(\xi_2)$ are initially unknown. The question is whether, from a given set of observations, we can infer with considerable posterior probability that m is finite, and that one or more of the observations have been affected by the error ξ_2. If so we are justified in rejecting them. Suppose then that we have n observations, that the largest residual is ξ, and that the standard error as computed from the whole of the observations is σ. Then if ξ is a fairly large multiple of σ the probability of getting one observation out of n in the range $d\xi$ is nearly $n(2\pi)^{-\frac{1}{2}}\sigma^{-1}\exp(-\xi^2/2\sigma^2)\,d\xi$ if the normal law and this value of σ are correct. Consider now the probability of an error in this range from some law other than the normal one. The aggregate of all such laws must be con-

sidered. It is plain that the chief contribution will come from those with m of the same order as n; for if m was much less than n we should expect a large fraction of the observations to be affected, while if m was much greater than n it would be unlikely that any would. Similarly the chief contribution will come from the values of σ_2 of the same order as ξ. For given m and σ_2 the probability of an error in a range $d\xi$ is of order $d\xi/4m\sigma_2$. The prior probabilities that m and σ_2 lie within the requisite ranges may be taken to be fractions, but not very small ones; let us say $\frac{1}{4}$. Then the prior probability of an error in the actual range, derived by way of such laws, is of order $\frac{1}{64}\frac{d\xi}{m\sigma_2}$ or of $\frac{1}{64}\frac{d\xi}{n\xi}$. The numerical coefficient is obviously capable of great variation. The prior probability of the normal law being of order unity, we can say that the ratio of the posterior probability that the error ξ_2 has contributed to ξ to the probability that it has not, is of order

$$\frac{1}{64n\xi}\frac{(2\pi)^{\frac{1}{2}}}{n}\sigma e^{\xi^2/2\sigma^2} = \frac{1}{30n^2}\frac{\sigma}{\xi}e^{\xi^2/2\sigma^2},$$

roughly. This gives a workable criterion. If this ratio is greater than 1, we may reject the observation; if it is less than 1, we should retain it. Otherwise, our observation may be rejected if $\frac{\sigma}{30\xi}e^{\xi^2/2\sigma^2}$ is greater than n^2. We have the following values:

$\frac{\xi}{\sigma}$	$\frac{\sigma}{30\xi}e^{\xi^2/2\sigma^2}$
1	0·03
2	0·12
3	1·00
4	25
5	1800

The question of rejecting an observation therefore does not arise unless ξ/σ is over 3; if we have 5 observations we may reject an observation with ξ/σ greater than 4; if we have 40

observations we may reject one with ξ/σ greater than 5; but then the function increases so rapidly that with any practicable number of observations we should reject those with ξ/σ greater than 6, and the inaccuracy of the coefficient $\frac{1}{30}$ is a matter of trivial importance.

A common astronomical practice is to reject automatically observations with residuals greater than 5 times the probable error, or 3·4 times the standard error, and to reject those with residuals greater than 3 times the probable error, or 2·0 times the standard error, if there is any intrinsic ground for doubting those particular observations. From the above considerations it appears that these rules are somewhat too stringent; 5 times and 3 times the standard error instead of the probable error would be better.

CHAPTER VI

PHYSICAL MAGNITUDES*

Multiplication is vexation;
Division's just as bad;
The Rule of Three perplexes me,
And Practice drives me mad.

NURSERY RHYME

6·1. The fundamental notion of any quantitative science is number. In its most elementary form this means the number of a class, and depends on the notion of the cardinal comparison of classes. Two classes of objects are said to be similar if their members can be arranged in pairs, one from each class, so that to every member of the one class corresponds one of the other, and none are left over. If such a correspondence is not possible the classes are not similar. Then any two similar classes have something in common, which is not shared by any class not similar to them. This property we call their *number*. All propositions about number are really propositions about the comparison of classes. In the works of Russell and Whitehead the definition is made apparently more precise by defining the number of a class as the class of all classes similar to the given class; this class being the same whatever one of the similar classes we begin with, all the classes have on this definition obviously the same number. But it might appear that on this definition the creation of a new class (a new set of ten things, for instance) makes some change in the class of all similar classes, and we cannot allow the number of a known class to be changed by such an event. Actually, however, Whitehead and Russell, in their *Principia Mathematica*, do not use this definition in practice; for they never explicitly use the notion of a class at all. They proceed by attaching a

* For a great many of the ideas in this chapter I am indebted to Dr N. R. Campbell's *Physics: The Elements*, though I do not agree with all he says. Cf. *Phil. Mag.* **46**, 1923, 1021–1025.

meaning to every proposition about the class, or the class of classes, which can be understood in terms entirely of more elementary ideas, but a class as such is never defined. From a physical point of view there seems to be no harm in supposing directly that classes exist and that similar classes have a common property, which we call their number. The advantages of the method of Whitehead and Russell are that it makes it possible to give a meaning to any proposition about numbers whether classes actually exist or not, and that it avoids the logical difficulties associated with the theory of types; but for our purposes these appear to be unnecessary refinements*.

From the notion of number we can proceed to those of the sum and product of two numbers. If two classes have no common member, and we form the class of the two together, the number of this class is called the sum of the numbers of the original classes. If we form the class of all possible pairs of members of the two classes, the number of this class is called the product of those of the original classes. If a class has no member its number is called 0. If when a is a member of a class any member of the class is identical with a, the number of the class is called 1. If a unit class is combined with a different unit class, the resulting class is said to have number 2, and so on. In this way the finite whole numbers can be defined, and their arithmetic can then be developed.

6·11. Number is an abstraction. When classes are similar in terms of our method of comparison of classes, member to member, we say that they have a common property, which we call their number. We say in fact that they have the *same* number, which is different from the number of any class not so comparable with them. If in whatever way the members of two classes are paired off there are always still some members of one left over when those of the other are

* Cf. Wittgenstein, *Tractatus Logico-Mathematicus*; F. P. Ramsey, *Proc. Lond. Math. Soc.* 25, 1926, 338–384.

exhausted, the class with the unpaired members is said to have the greater number, the other to have the smaller number. The observed fact is the result of the comparison; the property common to similar classes is an abstract idea derived from it. This derivation by abstraction is a logical step, and is of extremely wide application. We experience a similar sensation from the sight of blood, a brick, a sunset, and a Canadian apple; we abstract a common property, which we call redness, and which is not possessed by the midday sky, a lemon, or a tablecloth. All qualifying adjectives depend for their meaning on such processes, of different complexity in different cases. In such an expression as "ten men", "ten" is not an adjective qualifying "men"; this is seen at once if we try to attribute a meaning to "a ten man". "Ten" here qualifies a class of men; "ten men" really means "every man in a ten class of men". Sometimes, when objects are classified in terms of some method of comparison, the classes can be arranged in some definite order suggested by the method of comparison itself; thus we attach meanings not only to the statement that classes have the same number, but to the statement that one class has a greater or smaller number than another, and this makes it possible to arrange numbers in a definite order. This is the fundamental requirement of a physical magnitude. It is not possessed by all abstractions. For instance, we can classify objects according to the colour-sensation they give. But there is no direct reason suggested by our method of comparing objects according to colour to indicate what should be the order of arrangement of red, yellow, and brown. For this reason colour is not a physical magnitude. In the case of the pitch of a note, we can say directly from sensation that one note is higher or lower than another, and all pitches can be arranged in a single series based on this comparison. Something more is needed, however, before we can *measure* pitch. The existence of an order is necessary to measurement, but other conditions must be satisfied before we can make a quantitative determination.

6·2. The quantities capable of being measured directly are called fundamental magnitudes. Their character can be shown by considering one of the most important, namely length. When two objects can be placed so that they are in contact at both ends, we find by experiment that calipers or compasses adjusted so that they fit one object will also fit the other. Objects can then be classified together if they fit the calipers when the latter are kept in the same adjustment. We abstract the common property, which we call the length of the objects. But the method of comparison by juxtaposition of the objects, either directly or by way of the calipers, suggests a way of arranging them in order. If the calipers have to be set to a greater angle to fit one object than another, we say that the first has the greater length; our method of comparison not only gives a meaning to length, but arranges different lengths in order, so that any length is greater than any that precedes it in the order and less than any that follows it. Thus length, so far, is on the same footing as the pitch of a note. But there is a difference.

Consider the method of construction of a millimetre scale. A long screw is fixed so that it can turn in a bearing with a screw thread inside it. Whatever part of the screw is within the bearing, it fits. Every turn of the screw fits any turn of the bearing. In terms of our method of comparison, every turn of either therefore has the same length. In the manufacture of the scale, it is arranged that whenever the screw advances through a complete turn a device attached to it rules a transverse line on the scale. The object whose ends are two consecutive scale-divisions is therefore compared directly with the turn of the screw, which is known to have always the same length. Hence by the very definition of length every interval between consecutive divisions on the scale has the same length. When we measure a length we place the ends of the object in contact with the scale, or we apply calipers to the ends and apply the calipers to the scale; and we count the scale-intervals between the ends. The statement that the

length of an object is 153 mm. means then that the object has the same length as the object formed by placing 153 scale-intervals end to end, all the intervals by construction having the same length as the turn of a certain standard screw. We see now the difference between a length and the pitch of a note. When we put two objects together end to end along a scale we get a new object determined by the extreme ends; we say that in terms of our method of measurement, merely by counting scale-intervals, the combined object has a measure equal to the sum of those of the separate objects. But if we sound two notes of different pitches we do not get a single note of a new pitch. If the notes are sounded together we get a chord; if they are sounded in succession they give two distinct notes.

We can now specify in what conditions a property can be a fundamental magnitude. It must be possible to construct a scale such that every interval of the scale is the same in respect of that property, the test of being the same being comparison with some definite standard by the process that enables us to recognize differences in that property. The intervals must be consecutive, and the object must be measured by counting the number of intervals that it overlaps. When this is done the measure of the property is a fundamental magnitude. It has the property that if two objects of measures x and y are placed consecutively, the measure between the extremes is $x + y$.

Length is a fundamental magnitude. Angle, as measured by a protractor graduated in degrees, is another, for each degree-interval is compared with a standard length in the construction of the instrument. Time, or rather the interval of time required for a given process, is another. It is measured by counting the swings of a pendulum or a balance wheel, which occur in a definite order, so that each has an immediate successor, and this order is recognized directly by sight or sound. If two different mechanisms once take the same time to perform an oscillation, they do so again when compared again. When two processes, started at the same instant, also

end at the same instant, we classify them together and abstract the common property of the interval of time taken. We measure this by counting the number of oscillations of a standard instrument, say a seconds' pendulum, the balance wheel of a watch, or a tuning-fork, that take place during either process. By its essential structure the interval is therefore a fundamental magnitude. It is important that the interval is independent of the actual instant when the process starts, just as a length measured on a scale as the number of intervals overlapped by the object is independent of the position on the scale of the end first placed in position. Time may also be measured in terms of the rotation of the earth; the interval taken by the earth to turn through a standard angle is taken as the step, and any interval is measured as the number of times the earth has turned through this angle during the process. Angle being a fundamental magnitude, interval of time as measured in terms of it is another.

Mass, as found from a balance, is another fundamental magnitude. The bodies we call our "weights" are constructed so that they all counterbalance the same body on the other pan; and we can recognize when a body more than counterbalances, or fails to counterbalance, a body on the other pan. We classify together bodies that counterbalance the same body, and abstract the property of *mass*. If then a body counterbalances the same body as is counterbalanced by n of our standard weights, we say that its mass is n in terms of these weights. The number n is obtained directly by counting, and is evidently a fundamental magnitude.

6·3. Every fundamental magnitude is measured in terms of a certain property of its own kind, which we call the step of the instrument. In the case of number the step is the number 1. In the case of length, it is the length of the interval on the scale, or ultimately that of a turn of the fixed screw thread. For time, it is the interval between instants when the pendulum, balance wheel, or tuning fork passes its equilibrium position. For mass, it is the mass of the standard

weight. In all cases but number itself the standard is to a large extent at our disposal. In the case of length, for instance, instead of using a millimetre scale we might use a scale depending on a different screw thread, giving a scale divided into tenths of inches. The numbers obtained by measuring the same object on the two scales are different; the standard therefore matters. But we can compare different standards. Thus we find that an object measured in terms of a millimetre scale overlaps 254 intervals; measured in terms of a tenth-inch scale it overlaps 100 intervals. If we like we can test one scale against the other directly. The length of 100 intervals on the tenth-inch scale is then *the same* as the length of 254 intervals on the millimetre scale; it is the common property revealed by the method of comparison. Now such stretches on a scale may be placed end to end, and by the additive property of fundamental magnitudes it follows that if an object has the same length as $100x$ intervals on a tenth-inch scale, where x is any whole number, it also has the same length as $254x$ intervals on a millimetre scale. If we consider an object that covers 10 intervals on a tenth-inch scale, we cannot say immediately that it will cover 25·4 intervals on a millimetre scale, because so far we have attached no meaning to fractions of a scale-interval. Strictly speaking, the measures of length we have considered arise when the object exactly stretches from one scale-division to another. We cannot say at once that an object covers 25·4 intervals; but we can say that it covers more than 25 and less than 26 intervals. This must be so, for ten such objects placed end to end will cover 254 intervals on the millimetre scale. If each covered 25 intervals all ten would cover just 250 intervals; if each covered 26 intervals, the ten would cover 260 intervals. The question therefore arises, when an object has not the same length as an exact number of intervals on the scale, can we assign to it a measure? Evidently we can, in two different ways. We can read to the nearest whole number of scale-intervals. In that case we have to say

that, while 10 tenth-inch intervals and 25 millimetre intervals have the same length, and lengths have the additive property, 100 tenth-inch intervals and 250 millimetre intervals have not the same length. There is an apparent inconsistency, which can be removed only by recognizing that we are not dealing with a logical process, but with a physical law. We must admit the principle that our measures are liable to errors, arising in this case from the finite step of the instrument. The measure "25 millimetre intervals" is an approximation to the true length, not the actual length. The additive property of lengths, in fact, is a physical law. So long as we are dealing with exact multiples of the scale-interval its truth is merely a matter of counting. But when we have recognized that every object has a length and that most objects do not in fact cover an exact number of scale-intervals, we have to choose between the additive law and the adoption of an exact number for a measure. The additive law being a simple one, we therefore retain it as expressing the relation that holds between the true values, as defined in the last chapter, and regard departures from it as errors. In the case just considered, the measure of 25 scale-intervals has an error. The measure of ten similar lengths together is the same as that of 254 millimetre intervals; we retain the additive law and say that, since there is a length in each case, its measure can only be 25·4 millimetre intervals. This is the true value. When the length of one object is given as that of 25 millimetre intervals, we say that it has an error; if there are, as here, other means of fixing the true value, we say that the error is − 0·4 interval. Length is not a mere matter of counting; fractions must be admitted.

We can proceed as follows in finding a length. Suppose that the object to be measured is placed repeatedly against the scale, so that in each application the first end comes where the second end was in the previous application. In this way we, effectively, construct a new scale. At the mth application the total length overlapped is greater than that of r and less than

that of s scale-intervals; that is, we classify the whole numbers into two divisions, such that m applications of the object cover more of the scale than the number of intervals given by a number in the first division, and less of it than the number of intervals given by any number in the second division. If then we are to retain the additive property of length, we must say that m times the length of the object is greater than that of any number of intervals in the first class, and less than that of any number of intervals in the second class. Therefore the measure of the length of the object must be greater than that of r/m intervals and less than that of s/m intervals; and the greatest value of r is less by 1 than the least value of s. The measure is therefore specified within a fraction $1/m$ of a scale-interval. By varying m we can then find a series of intervals, each of which must contain the true value; alternatively, we divide the rational fractions into two sets, such that the number in the measure exceeds all in the first set and falls short of any in the second set. The true measure may then be any value between the largest in the first set and the smallest in the second. If m could be indefinitely large in practice this procedure would specify a cut in the rational fractions and define a real number. Actually there is a limit to the length of the measuring scale, and a certain amount of arbitrariness survives. It might be true, as far as we can tell, that every length can be associated with a number of scale-intervals expressed by a rational number.

But this principle becomes untenable when we consider more complicated laws. We find for instance that the square of the hypotenuse of an isosceles right-angled triangle must be twice the square of either side. If the measure of the side can be associated with a rational number, the only number that can be associated with the hypotenuse is $\sqrt{2}$ times that number, and $\sqrt{2}$ times a rational fraction cannot be rational. We need more numbers than rational fractions to keep our laws formally true. But if we admit all real numbers there is no difficulty.

The true length of an object corresponds then to a real number of scale-intervals. Now suppose that the object is compared with two different scales. The associated numbers are l and l'. Another object is compared with the same scales, giving numbers m and m'. Then we must have

$$\frac{l}{l'} = \frac{m}{m'}.$$

For suppose that we place the objects in steps along the scales, the first being repeated p times and the second q times, and consider the length of the object specified by going from the last mark on the first new scale to the last on the second. This has a length, on the first scale, equal to $pl - qm$; if this is negative it means that we have to go backwards. On the second scale we get similarly $pl' - qm'$. But if l/m and l'/m' were unequal we could find such values of p and q that q/p would lie between them. Then $pl - qm$ and $pl' - qm'$ would have opposite signs and we should have to go in opposite directions in the two cases to get from the end of one derived scale to the end of the other. But the objects specified by repeating the first p times and the second q times have definite lengths; the greater length will be the greater length whatever scale is used. Hence $pl - qm$ and $pl' - qm'$ must always have the same sign. Therefore $l/m = l'/m'$, or $l/l' = m/m'$. The numbers associated with any length on two scales are in a fixed ratio depending on the scales and not on the object itself.

6·4. Starting from the sensory notion of comparison of objects by juxtaposition, we have obtained the notion of length as a property by abstraction, and have shown how any length may be associated with a real number in relation to a certain scale. We can now proceed to the notion of length as a *quantity*. As a property of an object it is identified by a statement of the form "the length of the given object, in comparison with a millimetre scale, is specified by the number x". We write this in the form "the length of the object is x millimetres". On the face of it this statement is an abbrevia-

tion, and can be understood only by reference to the previous one and to the whole of the foregoing discussion. We have nowhere said what we mean by "a millimetre" as a noun, much less what we mean by "x millimetres". We might mean "the length of one interval on a millimetre scale". But length is a property, and we do not know what we mean by multiplying a property by x. We might attempt to re-analyse the statement by saying that "x millimetres" has a structure analogous to that of "ten men". Then it would have to mean "a class of millimetres, whose number is x". But clearly a class of lengths is not the same thing as any single length, even in the case where x is a whole number; and if x is a fraction there is no such thing as a class of number x.

There seem to be two possible attitudes to the statement "the length of the given object is x millimetres". We can take it as simply an abbreviation; if so there is nothing further to be said. But we may consider that "a millimetre" is something that exists and can be freely multiplied by real numbers to give other things of the same kind as itself. If so, "length" in this statement is no longer a property of the object. x times a property cannot in any sense be a property of the same kind. "Length" is now a new concept, called a *quantity*. There is no logical necessity for the existence of quantities; but for practical convenience of statement they are useful. The fundamental postulate of the theory of quantities is:

If the measure of a quantity is x, and the quantity is multiplied by the number y, we obtain a new quantity of the same kind whose measure is xy.

On our first analysis this is equivalent to:

If a property of an object, in comparison with a certain scale, is associated with the number x, and the property of the scale-interval on the first scale, when compared with a second scale, is associated with the number y, then when the property of the object is compared with the second scale it is associated with the number xy.

Here the measure of an interval of the first scale in terms of the second is y, and we have obtained a measure of x intervals of the first scale as equivalent to xy intervals of the second.

This proposition is true for fundamental magnitudes in consequence of 6·3. The importance of the expression of it in terms of quantities may be illustrated by reference to length. Suppose an object is measured in terms of a tenth-inch scale and that the associated number is x. We express this by saying that "the length of the object is x tenths of an inch". We measure an interval on the tenth-inch scale in terms of a millimetre scale, and find that the ratio of the two associated numbers is y. This, by 6·3, is the same for all intervals, and in particular when the interval is the interval between consecutive divisions on the tenth-inch scale. Then the length of the object, in comparison with the millimetre scale, is associated with the number xy, and we express this in the language of quantity by saying that "the length of the object is xy millimetres". That is, the statements "the length of the object is x tenths of an inch" and "the length of the object is xy millimetres" are completely equivalent for all values of x. In any proposition containing the expression "tenths of an inch" we can therefore replace every tenth of an inch by "y millimetres" without affecting the truth of the proposition. In this language, therefore, a tenth of an inch and y millimetres are completely equivalent ideas, and we can say

$$\text{a tenth of an inch} = y \text{ millimetres}.$$

It is this proposition that provides the usual rule for conversion of units from one scale of measurement to another.

Similar considerations apply to any fundamental magnitude. In the case of mass the actual boxes of weights used introduce a slight complication. We do not in practice weigh, for instance, in terms of milligram weights alone. We use weights found by experiment to be equivalent, in regard to objects counterpoised by them, to various multiples of the unit.

The process is equivalent to measuring a length in terms of decimetres, centimetres, and millimetres and using the known standards of comparison of the various units to reduce the whole to millimetres.

6·5. The majority of physical magnitudes are not measured directly. They occur as factors of proportionality in laws. Probably the only laws that do not involve such factors are those of simple constancy, and those expressing addition of measures of fundamental magnitudes in terms of the same scale. Nevertheless they may be connected with properties. For instance, liquids may be classified according to whether a given solid sinks or floats in them; and this relation is unaffected by the size and shape of the containing vessel, so long as the solid does not actually touch the sides. Using different solids we can classify liquids in terms of each. This method establishes an order among liquids and solids. It is found that if we have made the classification in terms of one solid and then try another, the latter either sinks in all the liquids that the first sinks in, or floats in all those that the first floats in. There may be an intermediate group such that one solid floats in them but the other sinks. The liquids may therefore be arranged in an order such that each supports all solids supported by those before it in the series, but will not support some solids supported by liquids after it in the series. Then each liquid is said to have a greater density than those that precede it and a smaller one than those that follow it. We have abstracted from the empirical relation the property of density. The process resembles in outline that of abstracting the notion of length from the results of juxtaposition. But the analogy breaks down at the next stage. We cannot construct a scale of comparison for density by combining objects. In dealing with length we could put two millimetre intervals in succession and call the length of an object that fits the two together 2 millimetres; in dealing with time a process such that a seconds' pendulum swings twice during it is said to

occupy two seconds. In each case two of the standard intervals together are greater, in terms of the method of comparison, than either separately. It is this fact that makes it possible to construct a scale. But in the case of density, when we put together two of the solids used for comparison, the combined solid does not determine a cut in the series of liquids outside those determined by the two solids separately; in general the cut it gives lies between the two former ones. There is no way of constructing a scale based on a single solid; and the measurement of density as a fundamental magnitude breaks down. But we can weigh a portion of a liquid on a balance, and find its volume by means of a measuring glass. Both volume and mass are fundamental magnitudes, and when the process is carried out several times on different portions of the same liquid it is found that they are in fact proportional; therefore they are connected by a differential equation of the form

$$dy/dx = y/x. \qquad (1)$$

This is a very simple equation, and its truth can be established with practical certainty by a very few trials. Its solution is

$$y = Ax, \qquad (2)$$

where A is what is known in pure mathematics as an arbitrary constant. What actually happens is that the integrated form is the first to be verified, and A is determined in the process of verification. But the actual observed values do not fit the form (2) exactly, but approximately. Nevertheless, since the law (2) is equivalent to the simple differential equation (1) we say that the equation represents a physical law, expressing the relation between volume and mass in portions of the same liquid. The arbitrariness in the solution is found to correspond to the differences between different liquids; all give the form (1), but in (2) the quantity A has different values for different liquids and therefore expresses a property of the liquid. We can, that is to say, arrange liquids

according to the values of A they give, and then give a name to A. It is then found that the order of increasing values of A is also the order specified by the results of flotation experiments. We then have a quantitatively determined value, the mass per unit volume, such that greater mass per unit volume among liquids corresponds completely to greater or less density. In this way we can attach a numerical value to density.

Density is an example of a *derived* magnitude. It is a property capable of being ordered, but not directly measured in terms of a single scale. A series of experiments must be conducted, such that in each experiment two fundamental magnitudes are measured; and the measures are found to be connected by a simple differential equation. This is then taken as the physical law. An adjustable constant emerges in the solution, and we call this constant the density. In general it appears that derived magnitudes are the adjustable constants that arise in the solution of the differential equations of physics. In the simple case of the comparison of two scales of measurement we have already introduced a derived magnitude, by saying that the length of an object is 2·54 mm. for every tenth of an inch. Here we have begun by establishing a rule of proportionality valid for any two scales, and have found the number 2·54 as the actual one applicable to the particular pair of scales chosen. But its character is less evident than in the case of density because the properties it enables us to compare are merely two different ways of specifying the same thing, the length of a given object. In the case of density the derived magnitude provides a means of connecting two quite distinct properties of a portion of the liquid, namely its volume and its mass.

In some sense a derived magnitude is measured in terms of a number, but its structure is more complicated than that of a fundamental magnitude. The units used in determining the various fundamental magnitudes involved are obviously reflected in the number that appears in the measure of the

derived magnitude. The number attached to a density as a mass per unit volume will depend on whether the mass is measured in grams or pounds, and the volume in cubic centimetres or cubic inches. When we say that a density is 1·34 grams per cubic centimetre, the expression "1·34 grams per cubic centimetre" is a complete entity; no item in it, neither "1·34", nor "grams", nor "cubic centimetre", can be changed without altering the meaning of the whole. For this reason it is incorrect to speak, as is done in many writings on the theory of dimensions, of a "mere change of units". There is no such thing as a mere change of units. If we alter a unit without altering the number in the measure, we are speaking of a different physical system, and cannot assert anything about it without a physical law to guide us; while if we already know the law a change of units tells us nothing that we cannot find out by keeping the same units and altering the numerical measure*.

In discussing length we began with length as a property of an object and led up to the idea of length as a quantity. Between the two a one-one correspondence exists. If two objects are different in the property, as tested by direct juxtaposition, they have different measures, and conversely. Similarly for density, we may regard it as a property, differences in which are tested by the method of flotation, or as a mass per unit volume, the mass and the volume being both measured as fundamental magnitudes. A one-one correspondence exists between the property and the measure. If we are to proceed to consider density as a quantity we must verify that its measure satisfies our fundamental law for quantities. To do this, consider a given portion of a substance, and measure both its mass and its volume in terms of two different scales. Suppose the numbers associated with the mass on the two scales to be m and m', and those associated with the volume v and v'. Then m'/m is μ, the measure of the interval

* For this reason the so-called "method of dimensions" is fallacious. It should be replaced by that of similarity, as Campbell has explained (*loc. cit.*).

of the first mass-scale in terms of the second, since mass is a fundamental magnitude; and v'/v is f, the measure of the interval of the first volume-scale in terms of the second, since volume is a fundamental magnitude. Hence

$$\frac{m'}{v'} = \frac{\mu}{f}\frac{m}{v}.$$

But m/v and m'/v' are the numbers associated with the densities on the two pairs of scales. If the density of a substance on the first pair of scales is associated with the number unity, then on the second pair it is associated with the number μ/f. Our equation enables us to extend this by saying that if the numbers associated with the density on the two pairs of scales are ρ and ρ', then ρ'/ρ, for all values of ρ, is the number associated on the second pair of scales with the density of a substance associated on the first pair of scales with the number unity. This shows that density actually does satisfy the rule required.

We saw that in any proposition about length we could replace a tenth of an inch by 2·54 mm. without affecting its truth or falsehood. Thus "x tenths of an inch" and "$2{\cdot}54x$ millimetres" express the same length, whatever x may be. Now when the scales are specified "ρ units of mass per unit of volume" expresses a definite density. Consider then a portion of the substance with a volume expressed by v and a mass expressed by ρv on the first pair of scales. We can say that its volume *is* v of the first volume-units, and its mass ρv of the first mass-units, using now the language of quantity. Also its volume *is* fv of the second volume-units, and its mass *is* $\mu\rho v$ of the second mass-units. We therefore have, for all values of v, the result that "ρv of the first mass-units per v of the first volume-units" expresses the density of the *same* portion of substance as "$\mu\rho v$ of the second mass-units per fv of the second volume-units". But since mass and volume in the same substance are proportional this implies that a density expressed by "ρ of the first mass-units per first volume-unit"

is the *same* as one expressed by "$\mu\rho/f$ of the second mass-units per second volume-unit". We can therefore replace any density ρ in terms of the first pair of scales by $\mu\rho/f$ in terms of the second pair. We can also, if we like, regard density now as the *ratio* of a mass to a volume. For if we choose to introduce the concept of the ratio of two quantities neither of which is a number, we can write the following equations:

$$v \text{ first volume-units} = fv \text{ second volume-units},$$
$$\rho v \text{ first mass-units} = \mu\rho v \text{ second mass-units}.$$

Hence by division

$$\frac{\rho v \text{ first mass-units}}{v \text{ first volume-units}} = \frac{\mu\rho v \text{ second mass-units}}{fv \text{ second volume-units}}.$$

But the constancy of the ratio of the numerical measures of mass and volume in the same substance entitles us to cancel the factor v in both ratios. Also if we call "one mass-unit per unit of volume" a "unit of density", we have

$$\rho \text{ first density-units} = \mu\rho/f \text{ second density-units},$$

which gives the correct law of conversion from the first pair of scales to the second. The notion of the ratio of two quantities of different kinds, though it resembles that of quantity itself in having no logical reason for its existence, can actually be shown to lead to correct answers, and is therefore justifiable on the ground of convenience. Every proposition containing it can if desired be reinterpreted in terms of more fundamental ideas and then verified.

Derived magnitudes, in comparison with fundamental ones, are less immediately related to sensation, but more general in application. Thus a specimen of a given liquid may have any mass or volume, but each of these fundamental magnitudes is directly determinable in terms of a standard and a definite method of comparison. Their ratio, however, is always the same (with precautions, if necessary, about keeping the temperature and pressure constant). The density is not directly measured, but remains the same for the same liquid however

we may vary the volume. Its very existence depends on the truth of the physical law that the ratio of the measures of the mass and the volume is constant, and therefore on the truth of the differential equation

$$\frac{d}{dv}\left(\frac{m}{v}\right) = 0,$$

which holds for any liquid. The constant ρ that occurs in the solution of this equation, namely

$$m = \rho v,$$

exists in consequence of the differential equation, which contains no quantity not measured directly; ρ is independent of the volume and therefore is more general in its application than the original data, but there is nothing in the work to indicate that it should be the same for different liquids, and it is in fact found to be different for different liquids.

6·6. Let us return to the question of the solid of revolution rolling down an inclined plane. It was found that the displacement was proportional to the square of the time, satisfying the equation

$$x = 0{\cdot}20t^2, \tag{1}$$

where x is measured in centimetres and t in seconds. The coefficient 0·2 is a ratio found by experiment to fit a number of observations and therefore represents a derived magnitude. In consequence of our earlier discussion of the probability of physical laws we cannot admit that a numerical constant in a law, in its ultimate form, is capable of continuous variation. But we can remove the constant by writing the law in any of the differential forms

$$\frac{d^3x}{dt^3} = 0; \quad \frac{d}{dt}\left(\frac{x}{t^2}\right) = 0; \quad \frac{dx}{dt} = \frac{2x}{t}, \tag{2}$$

of which the last two are equivalent. The first has the general solution

$$x = a + ut + \tfrac{1}{2}ft^2, \tag{3}$$

where a, u, and f are adjustable constants. The second and third have as their most general solution

$$x = \tfrac{1}{2}ft^2 \tag{4}$$

simply. In either case the constant $\frac{1}{2}f$ appears, and can be identified with the 0·20 of the actual experiment. But a and u are on a somewhat different footing. They resemble f in being constants of integration. But they are much more sensitive to the given experimental conditions. The form (4) is applicable only if x is measured from the initial position of rest and t from the time when the body is released. If the body is originally some way down the scale and moving, or if the stop-watch is not originally at zero, the form (4) is experimentally untrue, but (3) still holds with suitable values of a and u. But f is much more general in its application. Wherever the body is started we get the same value of f; a and u are more general than x and t, which vary from one single observation to another, but they vary from one experiment (series of observations) to another.

Actually we may repeat the experiment with a different inclination of the plane to the horizontal. It is then found that f itself is different, and is proportional to the sine of the inclination α. If we call this sine σ our law takes the simple form

$$\frac{df}{d\sigma} = \frac{f}{\sigma}. \tag{5}$$

But we may proceed to experiment with different solids of revolution, and we find that different solids give different values of f for the same inclination; in fact f is proportional to $c^2/(c^2 + k^2)$, where c is the distance from the axis of the body to the line of contact and k the radius of gyration about the axis. Both c and k can be found by measurement. The result of the several series of experiments is that the quantity

$$\frac{c^2 + k^2}{c^2} \frac{1}{\sin \alpha} \frac{d^2x}{dt^2} = g \tag{6}$$

is constant for variations of the initial position and velocity, of the inclination, and of the form of the section of the body by a plane through the axis. It is therefore a quantity of much greater generality than the actual acceleration in any one experiment, and its existence really sums up at once not one but several differential equations. This, then, is the ultimate form of the law of the rolling of a solid of revolution, and the constant g in it is the most general derived magnitude obtainable from such experiments. It can be shown by other experiments to be the acceleration of a falling body and to be also a derived magnitude associated with the simple pendulum. These agreements are predictable from the laws of dynamics and constitute a verification of these laws.

The quantities a and u are really derived magnitudes because they still arise in the solution of the general equation (6) and have to be determined in each experiment so as to make the formal solution fit the observed values of x and t as closely as possible. But in each experiment they are at our disposal; when one experiment is finished the values of a and u associated with it have no further application. Consequently they are not given in books of physical tables; but g is always given. There are cases, however, where a and u or their analogues have a wide application. We do record the position of a star and its proper motion at the date 1900.0. The reasons are first, that these quantities are useful in finding the position of the star at any other date, years or centuries earlier or later; and second, that we cannot put the star back and start it off differently, so that the quantities are not at our disposal. In fact so far as experiments go the difference in character between a and u, f, and g is one of degree of generality; there is no fundamental qualitative difference.

6·7. In all numerical work we make free use of mathematics; and the numbers that arise go far beyond the simple class-numbers that we start with. We have already seen, in considering the properties of the isosceles right-angled triangle,

that we cannot even restrict the values of our quantities to rational multiples of the unit without sacrificing the truth of the laws of physics. The adopted values of the lengths of the sides, if obtained by making repeated measures to the nearest multiple of the scale-interval and taking the mean, will always be rational. We can preserve the exactness of the law only by admitting the existence of errors in the adopted values themselves; and that implies the existence of an unknown *true value* behind the adopted value. This is the justification of the assumption we made in discussing errors of measurement, that there is a true value, which our observations may enable us to identify within limits, but never exactly. Further, the whole series of rational numbers is insufficient to specify all the possible true values; and there seems to be no reason for not admitting the whole series of real numbers.

This brings us to an attitude towards real numbers that seems to agree better with that of ordinary mathematicians than with that of Whitehead and Russell. The usual procedure in defining $\sqrt{2}$, for instance, would be to divide the rational fractions into two classes, such that the squares of those in one class were all less than 2, and of those in the other class greater than 2. The number separating the two classes is then called $\sqrt{2}$. The principle is known as Dedekind's section. Whitehead and Russell, however, point out that there is no *a priori* reason to believe that there *is* any number that is greater than all numbers of the first class and less than all of the second class; Dedekind's section assumes an existence without proof. Whitehead and Russell proceed by defining $\sqrt{2}$ as the class of all rational fractions whose squares are less than 2; this class exists in the same sense as other classes. But addition and multiplication of classes have to be redefined for non-rational numbers so as to keep the ordinary laws of algebra true. Thus $\sqrt{2} + \sqrt{3}$ has to be defined as the class of the sums of all pairs of rational fractions such that one has a square less than 2 and the other a square less than 3, and $\sqrt{2} \times \sqrt{3}$ as the class of the products of pairs of rational

fractions satisfying the same conditions. In this way they are able to establish the existence theorem for real numbers and develop their algebra without assumption. But from the physical point of view there is no apparent reason to believe that the numbers that occur in true values of variable quantities are really classes of rational fractions, while there is direct reason to believe that these numbers do exist and are different from rational fractions. From our point of view Dedekind's assumption is therefore open to less objection than that of Whitehead and Russell. The utility of the latter is that it does establish what might otherwise be open to a certain amount of doubt, that there is no internal inconsistency in assuming the existence of non-rational numbers and applying to them the ordinary rules of algebra established for rational numbers.

CHAPTER VII

MENSURATION

'Tis distance lends enchantment to the view.
THOMAS CAMPBELL, *Pleasures of Hope*

7·1. We are now in a position to begin the discussion of the most fundamental physical science, that of the relations between lengths. We shall call it *mensuration*. It requires to be distinguished at the outset from the subject known to pure mathematicians as *geometry*. The latter is a branch of pure logic. It proceeds by taking a number of general axioms, irrespective of whether these are physically tested or capable of being tested, and develops their consequences by purely logical rules. Physical measurement plays no part in it. For us, physical measurement is the whole *raison d'être* of the subject. By comparing our measurements we establish certain laws; these lead to generalizations, which in many cases resemble the axioms of forms of geometry. But the structure is essentially different. In geometry the laws are assumed *a priori*, and the particular results are consequences of the laws. In mensuration the particular results are the essence of the matter, and the general laws are derived from them by a process of generalization based on the simplicity postulate.

It might nevertheless appear that, in spite of the opposite modes of approach, the total content of mensuration and geometry might be the same, the axioms of geometry being the same as the laws of mensuration. But this is not the case.

All projective and descriptive geometries can evidently be ruled out at once. A requirement of all such geometries is that no notion analogous to distance is to be used. Since distance forms our subject-matter, there is no common ground whatever.

Euclid's geometry is the closest existing analogue of mensuration, and requires a full discussion. The notion of length is freely used in it. His points are sufficiently like what we have so far called the ends of objects, and we can produce close enough physical analogues of his straight lines, planes, and circles. He freely uses the principle of juxtaposition as a test of whether one quantity is greater or less than another; here he follows the ordinary physical method of comparing lengths.

Nevertheless his system differs from any possible system of mensuration; in fact it is neither a mensuration nor a geometry, but a mixture of the two. For instance, he uses compasses to draw circles, a legitimate physical procedure, but refuses to use them to transfer distances. In I (2), when he wishes to draw from a given point a line whose length is equal to that of a given straight line not through the point, he makes a complicated construction to avoid having to lift up the compasses and transport them. Yet in I (4), in testing the equality of two triangles, he supposes one picked up bodily and superposed on the other. The ordinary properties of rigid bodies are supposed to be possessed by triangles (drawn on pieces of paper) but not by a pair of compasses. The usual criticism from the geometrical standpoint is to reject the proof of I (4) and provide a new one; from the physical one the proof of I (4) is valid in certain conditions, though the result is true even when the construction involved cannot be carried out, but the complication of I (2) avoids only a difficulty that does not exist.

Euclid postulates further that any two points can be joined by a straight line. The physical analogue of this is often true, but not always. The points may be on the surface of a convex body too hard to be bored. Yet the distance between the points exists, for it can be measured by applying compasses first to the two points and then to a scale. Physically the notion of the distance between two points is more general than that of the straight line joining them.

The most important departure of Euclid's treatment from any possible account of mensuration, however, is in the discussion of parallels and the related propositions. We may refer to the second postulate, that a straight line can be produced to any length, however great, and to the fifth postulate, also called the twelfth or parallel axiom. Both of these postulates have been criticized by modern geometers as not obvious. In mensuration, on the other hand, they are not only not obvious but demonstrably false. We cannot produce a physical straight line to a length greater than one determined by the size of the body it is drawn on; it may be extended by fastening other bodies on, but there is a limit to this process, and therefore to the length of the line. Again, it may be possible to find out by our existing methods of measuring angles that, when one straight edge crosses two others, it makes the sum of the interior angles less than two right angles, but it does not happen in all such cases that the two straight edges it crosses intersect; for in practice they often cannot be made long enough, or they may not be in one plane —a detail not allowed for in the usual statement of the postulate.

The alternative known as Playfair's axiom does not meet the difficulty, for it is not true that of two intersecting straight edges at least one must intersect any other; Playfair's parallel axiom fails in just the same way as Euclid's.

Criticisms of Euclid's *Elements* have usually been made from the geometrical standpoint and not from the physical one, and its virtues from the one are usually its vices from the other. His test of equality is always superposition. Two lengths are equal if one can be superposed on the other. The same applies to two angles. Two areas are equal if one can be cut up and the pieces placed so that they exactly cover the other. These are physical methods of comparison; and just for that reason they are rejected by geometry. His procedure is such that numerical measures do not arise; addition and subtraction of quantities are done on the actual objects them-

selves. He thereby sacrifices the convenience of being able to resort to algebra; but he also avoids a trap. Euclid would never have said that a length of 1·5 cm. is converted into one of 1·5 in. by a "mere change of units"; nor would he have said that the mass of the sun is 1·5 kilometres.

The word "geometry" literally means the measurement of the earth, and Euclid's predecessors were doubtless largely inspired by the needs of surveying. By this time, however, the name has become so closely connected with the branch of pure mathematics that it seems hopeless to rescue it. Nor is it, I think, worth while. The measurement of the earth is now generally known as "geodesy", and what we need is a word to describe the theory of measurement of length in general, not merely in relation to the earth. "Mensuration" seems entirely satisfactory, saying neither more nor less than it actually means.

It is a fact that when Euclid's theory gives a quantitative result, and the relevant construction can be carried out, the result is always found to be physically correct*. Nevertheless his axioms assume so many things possible that are in fact physically impossible that a radical reconstruction is needed. The modern physicist will not share his antipathy to numerical measurement, and will recognize in his treatment of angles and areas a perception that these, like length, are fundamental magnitudes. If he accepts the notion of quantity he will not refuse to say that a square centimetre is literally the square of a centimetre; but it is not strictly necessary to say so. The question that does actively arise at the outset, however, is whether we should introduce from the start any fundamental magnitudes besides length. Euclid assumes in I (13) that if a pencil of coplanar lines is drawn through a point, the angle between the extreme lines is equal to the sum of those between consecutive lines of the pencil; and in I (4) he supposes that angles that can be superposed are equal. These postulates

* Except in the extreme case of the displacement of star images by the sun's gravitational field.

make it possible to construct a scale for measuring angles in terms of a unit. Such a scale we may at once call a protractor, and angle is a fundamental magnitude. Again, in I (35) he compares the areas of parallelograms by cutting them up and superposing them, and his later work with triangles and rectangles indicates that area also is a fundamental magnitude. There are therefore three different fundamental magnitudes in the theory, and in the development they continually influence one another. All can be shown to exist in the sense that their measurement can actually be carried out, and there is no theoretical objection to developing the theory of all together. But there is a practical objection. Angles and areas can actually be superposed only in special cases; projections on the bodies that carry them usually interfere with the superposition. Again, the addition of angles or areas is meaningless unless they are placed in the same plane; thus the direct measurement of either depends on the existence of physical planes, whereas the measurement of distance by means of compasses and scale is independent of the existence of planes. Since distance is much more generally measurable directly than either angle or area it is desirable to develop the theory, if possible, on the basis of the properties of distance alone.

7·2. Mensuration deals essentially with the relations between measurements of distance on rigid bodies. It may be suggested that before it can be discussed we should define the terms "distance" and "rigid bodies". Now the requirement of a definition is that it must make it possible to recognize the defined object when it actually occurs. It is of no value to say that a rigid body is one such that the distances between all the points of it are unaltered by any displacement, nor to define relative motion as change of distance between parts of a system, unless we have some way of recognizing when distances *are* altered. Distance, again, cannot be defined in terms of the properties of rigid bodies unless we have first some way of recognizing the rigid body when we meet it.

None of these notions can be defined in terms of the properties of "space", because we have no means of recognizing space directly; distance in space, for instance, cannot be determined except through measurements, which at once reintroduce material scales, which the reference to space was intended to avoid.

The solution seems to be that neither "rigid body" nor "distance" is directly recognizable, and that both are derived from still more elementary notions, several experimental facts being used in the process. Let us start from the notion of a *body*, without considering how we arrive at this concept. It is a fact that we can make permanent marks on bodies, which we can recognize afterwards. It may be found that if we have two marks A, B on one body and two others, C, D on another, we can place the bodies so that A coincides with C and B with D. If a pair of compasses or calipers is adjusted so that one point coincides with A and the other with B, then it can be transported without readjustment and placed so that one point coincides with C and the other with D. All pairs of marks that can be fitted by the compasses in the same adjustment are classified together; we abstract the common property of *distance*, and say that all such pairs are equidistant. It may happen that two equidistant pairs can be superposed directly; but this is not always possible, because material obstructions may interfere. It is clear, in particular, that different pairs of marks on the same body cannot be superposed without deforming the body, even if they are equidistant. Now when a fit of pairs of marks has been obtained, either directly or through the use of compasses, it may be found that a fit is always obtained again in any subsequent trial. If this holds for numerous pairs of marks on the same body, we can generalize it as a law for that body. Such a body is called *rigid*. Compasses are rigid bodies provided their adjustment is not altered. If there is a doubt as to whether the adjustment has altered they can be tested by application to several pairs of marks that they previously

fitted; and if they fail we can tighten up the hinge or get a new pair. In the first place distance is simply a property of pairs of marks on rigid bodies.

It is also a fact that bodies can be made with *edges*; if two bodies touch at two points they may touch at a continuous set of intermediate points. In general, when this is done, if we turn one or both of the bodies about so that they remain in contact at two given pairs of marks, the intermediate marks that were formerly in contact separate. But it is again a fact that bodies can be made with such edges that they do remain in contact at intermediate marks when they are turned about two coincident marks. When this has been found to hold in a number of trials it can be inferred with a high degree of probability that it will hold in any subsequent trial. In such cases we call the edges *straight*.

The reservation must be made that the bodies, in both types of test, must receive only ordinary treatment. It is easy to recognize by the sensations that we call sensations of force when exceptional treatment is taking place. If bodies or edges fail to satisfy our tests we say that they are not rigid or not straight, or that exceptional treatment has taken place. In that case they do not form part of our present subject-matter. The important thing is that there are many bodies that do satisfy the conditions. If all compasses were made of rubber and all bodies of plasticene, this would not be so, and then perhaps there would be no science of mensuration; but actually we can classify bodies and edges according as they do or do not satisfy our tests, and confine our attention for the present to those that do. The others are reserved for the subjects of mechanics.

So far we have been able to define only *identity* of distance and not the meaning of *greater* and *less* in relation to distance. We need also to be able to establish a meaning for the statement that the distance AB between one pair of marks is greater than the distance CD between another pair. To do this it is necessary to be able to establish two other pairs of

marks A' and B', C' and D', such that by our criterion the distance AB is the same as the distance $A'B'$, and the distance CD the same as the distance $C'D'$, and so that $A'B'$ and $C'D'$ are directly comparable. One method of comparison would be to say that $A'B'$ is greater than $C'D'$ if the compasses have to be set to a wider adjustment to fit the former. Another, which is more closely related to actual measurement, is to use the straight edge. If A' and B' are on a straight edge, there is a definite path from A' to B' along the edge. If the marks C' and D' are chosen on the same edge so that one lies *between* A' and B' and the other either also lies between them or coincides with one of them, then the part of the edge $A'B'$ includes the whole of the part from C' to D' with something over, and we say that the distance $A'B'$ is greater than the distance $C'D'$, and therefore that the distance AB is greater than the distance CD. The introduction of the straight edge in defining the meaning of *greater than* in relation to distance seems to be necessary because, while unequal distances cannot in any case be superposed by moving the respective solids about, there are also cases where the distances are really equal but cannot be superposed on account of the form of the solids, and we must provide ourselves with a means of distinguishing between real difference and mere failure to carry out a strict comparison.

We can now proceed to the construction of a measuring scale on a straight edge and to the actual measurement of distance by the principles of the previous chapter.

7·3. So long as marks lie on the same straight edge, the distances between them follow the simple rules of addition and subtraction. But we also require propositions connecting distances between marks not on the same straight edge. This brings us into a new domain, and at least one new experimental fact is needed to serve as a starting-point. As has already been indicated, propositions involving angles or planes should be avoided as far as possible until we can define

them in terms of lengths. Our physical treatment must begin with an experimental law connecting distances between marks not on the same straight edge.

7·31. Consider any three marks, O, X, Y whose mutual distances are known, and consider the ratio

$$\lambda = \frac{OX^2 + OY^2 - XY^2}{2OX.OY}. \qquad (1)$$

If O, X, and Y lie on the same straight edge, and O is not between X and Y, then

$$XY = OX \sim OY, \quad \text{and} \quad \lambda = 1. \qquad (2)$$

If O, X, and Y lie on the same straight edge, and O is between X and Y, then

$$XY = OX + OY, \quad \text{and} \quad \lambda = -1. \qquad (3)$$

If O, X, and Y are not on the same straight edge, λ in general is found by experiment to lie between ± 1.

But if X and Y lie on two rigidly fixed straight edges meeting in O, then wherever X and Y may be taken on these edges λ has a constant value; that is, λ is independent of both OX and OY.

This proposition lacks the chief requirement of a postulate in a geometry, namely that of possessing a *naïveté* that disarms suspicion. But for our purpose what matters is that there should be a practical way of ascertaining whether it is true; and it is capable of test in almost all cases, and such test has already been carried out in countless experiments in practical plane "geometry". It has perhaps not been tested directly with the full accuracy of modern measuring apparatus, but enough has been done to establish it in an enormous number of cases. It is not extremely simple in form, but the number of verifications is so great that if it has any appreciable prior probability the probability of all inferences from it must amount to practical certainty. We therefore suppose it to hold in general and attempt to develop its consequences.

7·32. Instead of using the ratio λ itself, it is convenient to work with a certain function of it. We put

$$\lambda = \cos \alpha, \tag{4}$$

where the cosine is defined as in works on analysis. This defines a value of α less than π. Also since λ is independent of the actual values of OX and OY, its value expresses a property of the pair of edges OX and OY as wholes and not of any particular marks on them. The same applies therefore to α. We denote α usually by $\angle XOY$ and call it the *angle between OX and OY*. Then (1) is equivalent to

$$XY^2 = OX^2 + OY^2 - 2OX.OY \cos XOY. \tag{5}$$

This is practically Euclid II, 12 and 13.

7·33. It may happen that $\lambda = \pm 1$ when the three marks do not lie on a straight edge. In any case when $\lambda = \pm 1$ we call the marks *collinear*. If $\lambda = -1$ we say that O is between X and Y, and $\angle XOY = \pi$; if $\lambda = +1$, we say that O is not between X and Y, and $\angle XOY = 0$. It is also found that if O, X_1, X_2 are collinear, and O, Y_1, Y_2 are collinear,

$$\frac{OX_1^2 + OY_1^2 - X_1Y_1^2}{2OX_1.OY_1} = \frac{OX_2^2 + OY_2^2 - X_2Y_2^2}{2OX_2.OY_2},$$

irrespective of whether the collinearities are given by actual straight edges. Also there is only one possible position Y collinear with O and X, such that OY has a given value and O is not between X and Y. Conversely in experiments on a laboratory scale, if O and X are already assigned, we can place Y so that OY has any given value. We can also generalize to collinear sets of points the principle corresponding to Euclid's postulate that two lines cannot enclose a space, namely that as Y proceeds from O to X and beyond it there is only one possible path such that the points O, X, and Y are always collinear and along this OY increases continuously.

We can now proceed to develop the theory*.

7·41. If A, B, C are three marks as in Fig. 1, we have

$$\cos BAC = \frac{AB^2 + AC^2 - BC^2}{2AB.AC}.$$

Write $BC = a$, $CA = b$, $AB = c$, and $2s = a + b + c$. Then

$$\cos BAC = \frac{b^2 + c^2 - a^2}{2bc},$$

$$\sin \tfrac{1}{2}BAC = \left(\frac{1 - \cos BAC}{2}\right)^{\frac{1}{2}} = \left(\frac{-(b-c)^2 + a^2}{4bc}\right)^{\frac{1}{2}} = \left\{\frac{(s-b)(s-c)}{bc}\right\}^{\frac{1}{2}},$$

$$\cos \tfrac{1}{2}BAC = \left(\frac{1 + \cos BAC}{2}\right)^{\frac{1}{2}} = \left(\frac{(b+c)^2 - a^2}{4bc}\right)^{\frac{1}{2}} = \left\{\frac{s(s-a)}{bc}\right\}^{\frac{1}{2}},$$

$$\tan \tfrac{1}{2}BAC = \frac{\sin \tfrac{1}{2}BAC}{\cos \tfrac{1}{2}BAC} = \left\{\frac{(s-b)(s-c)}{s(s-a)}\right\}^{\frac{1}{2}}.$$

Corresponding formulae for the other angles are obtained symmetrically.

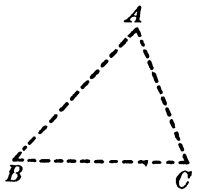

Fig. 1

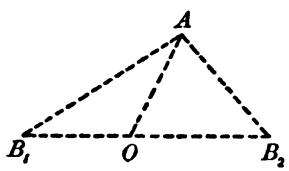

Fig. 2

7·42. By the formula for the tangent of the sum of two angles,

$$\tan \tfrac{1}{2}(BAC + BCA) = \frac{\tan \tfrac{1}{2}BAC + \tan \tfrac{1}{2}BCA}{1 - \tan \tfrac{1}{2}BAC \tan \tfrac{1}{2}BCA} = \left\{\frac{s(s-b)}{(s-c)(s-a)}\right\}^{\frac{1}{2}} = \tan(\tfrac{1}{2}\pi - \tfrac{1}{2}ABC).$$

Hence $\angle BAC + \angle ABC + \angle BCA = \pi$. Euc. I, 32.

* In the figures continuous lines denote actual straight edges; dotted lines connect only marks the distances between which are considered, but which need not be connected by actual straight edges.

7.43. If B_1, O, B_2 are collinear marks (Fig. 2), and A is another mark,

$$AB_1^2 = OA^2 + OB_1^2 - 2OA.OB_1 \cos AOB_1, \quad (1)$$

$$AB_2^2 = OA^2 + OB_2^2 - 2OA.OB_2 \cos AOB_2, \quad (2)$$

and also $$= AB_1^2 + B_1B_2^2 - 2AB_1.B_1B_2 \cos AB_1B_2. \quad (3)$$

But $$B_1B_2 = B_1O + OB_2, \quad (4)$$

$$\cos AB_1B_2 = \cos AB_1O = \frac{AB_1^2 + OB_1^2 - AO^2}{2AB_1.OB_1}$$

$$= \frac{OB_1 - OA \cos AOB_1}{AB_1}, \quad (5)$$

by (1). Substituting from (4) and (5) in (3) we have

$$AB_2^2 = AB_1^2 + B_1B_2^2 - 2B_1B_2\,(OB_1 - OA \cos AOB_1)$$

$$= OA^2 + OB_1^2 + B_1B_2^2 - 2B_1B_2.OB_1 + 2OA\,(B_1B_2 - OB_1) \cos AOB_1$$

$$= OA^2 + OB_2^2 + 2OA.OB_2 \cos AOB_1, \quad (6)$$

whence, comparing (2) and (6),

$$\cos AOB_1 + \cos AOB_2 = 0, \quad (7)$$

and therefore $$\angle AOB_1 + \angle AOB_2 = \pi. \quad (8)$$ Euc. I, 13

7·44. It follows as an immediate corollary by Euclid's method that when two straight edges cross the opposite angles are equal. Euc. I, 15.

7·45. It also follows from 7·42 that

$$\angle AB_1O + \angle B_1OA + \angle OAB_1 = \pi, \quad (1)$$

$$\angle OB_2A + \angle B_2AO + \angle AOB_2 = \pi, \quad (2)$$

$$\angle B_1B_2A + \angle B_2AB_1 + \angle AB_1B_2 = \pi. \quad (3)$$

Adding (1) and (2) and subtracting (3), and cancelling identical angles,

$$\angle B_1OA + \angle AOB_2 + \angle OAB_1 + \angle B_2AO - \angle B_2AB_1 = \pi. \quad (4)$$

But by 7·43 $$\angle B_1OA + \angle AOB_2 = \pi. \quad (5)$$

Hence $$\angle OAB_1 + \angle B_2AO = \angle B_2AB_1. \quad (6)$$

Thus angles so placed that they have one common arm and so that collinear points exist on the arms, and measured by the rule 7·32 (4), have the additive property.

7·46. If two straight edges OA and OB meet in O, and if $\cos AOB$ is negative, and if O is not the end of OB, it is possible to make a mark B_1 on OB so that $\cos AOB_1$ is positive.

For by 7·43 we need only take B_1 on the side of O opposite to B, and the result follows.

7·47. If A be outside the edge OB, and if B be on that part of it where $\cos AOB$ is positive, and if OB is greater than $OA \cos AOB$, then it is possible to make a mark C on OB such that

$$OA^2 = OC^2 + AC^2.$$

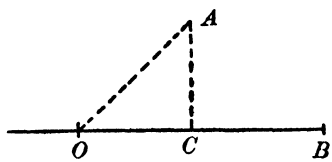

Fig. 3

For we can make a mark C on OB at a distance $OA \cos AOB$ from O. Then

$$AC^2 = OA^2 + OC^2 - 2OA.OC \cos AOB$$
$$= OA^2 - OC^2, \quad (1)$$

which proves the proposition.

Also
$$\cos OCA = \frac{OC^2 + CA^2 - OA^2}{2OC.CA} = 0, \quad (2)$$

whence
$$\angle OCA = \tfrac{1}{2}\pi. \quad (3)$$

We have therefore constructed a triangle with one angle equal to $\frac{1}{2}\pi$. We can now introduce the definition of perpendicularity. If the angle between two intersecting straight edges is $\frac{1}{2}\pi$, they are said to be *perpendicular*. Euclid I, 47 follows immediately from the formula 7·32 (5).

If OB is a straight edge with a mark C on it such that $\angle OCA$ is $\frac{1}{2}\pi$, where A is a mark not on OB, then C is called the foot of the perpendicular from A to OB.

7·48. If two straight edges OA, OB intersect at O (Fig. 4), and C is any mark in OA, and if the length of OB exceeds

OC sec AOB, then we can make a mark D on OB so that C is the foot of the perpendicular from D to OA.

For we can make a mark D so that $OD = OC$ sec AOB, and the perpendicularity follows as in 7·47.

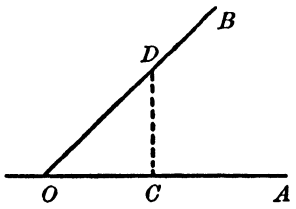

Fig. 4

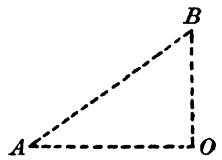

Fig. 5

7·49. If the angle AOB (Fig. 5) is $\frac{1}{2}\pi$, we have

$$\cos OAB = \frac{OA^2 + AB^2 - OB^2}{2OA.AB} = \frac{OA}{AB}, \qquad (1)$$

$$\sin OAB = \cos(\tfrac{1}{2}\pi - OAB) = \cos OBA \text{ by } 7\cdot42$$
$$= \frac{OB}{AB}, \qquad (2)$$

$$\tan OAB = \frac{\sin AOB}{\cos AOB} = \frac{OB}{OA}, \qquad (3)$$

with corresponding formulae for the other trigonometric functions. These results thus emerge as laws, and not as definitions of the functions as in ordinary trigonometry.

7·50. Consider three edges meeting in a point O (Fig. 6). It is always possible to fix A in one of them so that A is the foot of the perpendiculars from marks B and C on the other two, since the condition of 7·47 can always be satisfied by making OA short enough. Then

$$AB = OA \tan AOB, \qquad (1)$$

$$OB = OA \sec AOB, \qquad (2)$$

$$AC = OA \tan AOC, \qquad (3)$$

$$OC = OA \sec AOC, \qquad (4)$$

$$BC^2 = OB^2 + OC^2 - 2OB.OC \cos BOC$$
$$= OA^2(\sec^2 AOB + \sec^2 AOC$$
$$- 2 \sec AOB \sec AOC \cos BOC). \qquad (5)$$

Also

$$BC^2 = AB^2 + AC^2 - 2AB.AC \cos BAC$$
$$= OA^2 (\tan^2 AOB + \tan^2 AOC - 2 \tan AOB \tan AOC \cos BAC). \quad (6)$$

Equating (5) and (6), and multiplying by $\cos AOB \cos AOC$, we have

$$\cos BOC = \cos AOB \cos AOC + \sin AOB \sin AOC \cos BAC. \quad (7)$$

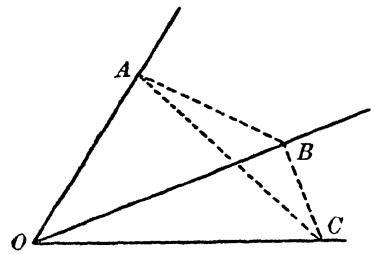

Fig. 6

This theorem introduces the third dimension for the first time, for it allows two different lines AB, AC to be both perpendicular to an edge OA at the same point. The formula (7) is the analogue of a familiar one in spherical trigonometry, though the sphere as such has not yet appeared.

It follows as a corollary that $\angle BAC$ is independent of OA.

If B, A, C are collinear, $\cos BAC = -1$, and (7) leads to

$$\angle BOC = \angle AOB + \angle AOC.$$

This is equivalent to 7·45 when actual straight edges connect the marks.

7·51. We can now proceed to a discussion of *planes*. If we take two fixed marks O and O', and any path from O to O', then at one end of the path the distance from O is greater than that from O', and at the other end the opposite is true. Both distances vary continuously, and

therefore there is a possible position on the path such that the distances from O and O' are equal. Marks equidistant from two fixed marks are said to be *on the plane* determined by the fixed marks.

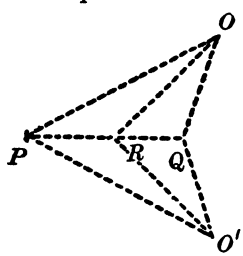

Fig. 7

7·52. If two marks P and Q are on a plane, every mark R collinear with them is on the plane. For since P and Q are on the plane we can write

$$PO = PO' = p; \quad QO = QO' = q; \quad PQ = r, \quad (1)$$

and

$$\cos OPQ = \cos O'PQ = \frac{p^2 + r^2 - q^2}{2pr} = k, \text{ say.} \quad (2)$$

Then
$$\begin{aligned} RO^2 &= PO^2 + PR^2 - 2PO.PR \cos OPQ \\ &= p^2 + PR^2 - 2p.PR.k \\ &= RO'^2 \text{ by symmetry.} \end{aligned} \quad (3)$$

7·53. If two planes are determined by pairs of marks O and O', H and H', three circumstances may arise. All positions P in the first plane may be equidistant from H and H'; then the planes are identical. All positions P in the first plane may be such that $PH > PH'$, or all such that $PH < PH'$: then the planes have no common point. Some positions P in the first plane may be such that $PH > PH'$, and others such that $PH < PH'$. Then we may classify the positions on the first plane according to the sign of $PH - PH'$; on any path from a position where this is positive to one where it is negative, the difference varies continuously and therefore passes through the value zero. Thus it is possible to assign marks common to both planes; they are said *to be on the line of intersection of the planes.*

All marks on the line of intersection of two planes are collinear. For if Q and R are on this line, every mark collinear with Q and R is common to both planes, by 7·52; hence the marks collinear with Q and R constitute the line of intersection. Euc. XI, 3.

7·54. Any three marks A, B, C lie on a plane. For if we take a point B' collinear with BA so that $B'A = AB$, and one C' on CA so that $C'A = AC$, then B' and B determine one plane and C' and C another. Clearly A lies on both planes. If O is another common point (Fig. 8) we have

$$OB' = OB; \quad OA = OA; \quad AB' = AB,$$

and therefore $\angle B'AO = \angle OAB = \frac{1}{2}\pi$.

Thus AO is perpendicular to AB, and similarly to AC. Now take O' collinear with OA so that $AO' = OA$. Then

$$\angle OAB = \angle O'AB = \tfrac{1}{2}\pi; \quad AO = O'A; \quad AB = AB;$$

and therefore $O'B = OB$.

Similarly $OC' = OC$,

and A, B, C are all on the plane determined by O and O'.

It follows from 7·54 and 7·52 that if we start with any three marks we can generate a plane containing them by joining up points collinear with pairs from the original three.

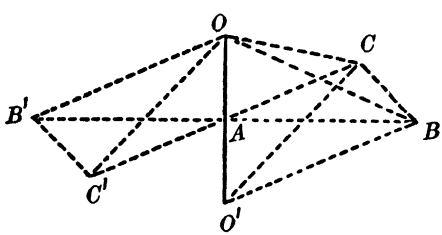

Fig. 8

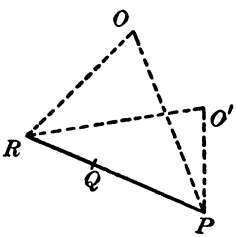

Fig. 9

7·55. In general a line has one point in common with a plane. For if P and Q are marks on the line (Fig. 9) and O, O' determine the plane, and R is another point on the line, in the direction PQ,

$$OR^2 = OP^2 + PR^2 - 2OP.PR\cos OPQ, \qquad (1)$$

$$O'R^2 = O'P^2 + PR^2 - 2O'P.PR\cos O'PQ, \qquad (2)$$

and therefore $OR = O'R$ if

$$2PR\,(O'P\cos O'PQ - OP\cos OPQ) = O'P^2 - OP^2. \qquad (3)$$

If then $O'P\cos O'PQ - OP\cos OPQ$ and $O'P^2 - OP^2$ have

the same sign, there is a positive value of PR satisfying (3). If they have opposite signs there is no positive value of PR satisfying (3). But if Q' is on the line and P is between Q and Q', there is a suitable point R in the direction PQ', since

$$O'P\cos O'PQ' - OP\cos OPQ' = -(O'P\cos O'PQ - OP\cos OPQ).$$

Clearly the conditions for suitable positions of R in the directions PQ and PQ' are mutually exclusive, and there is always one position of R that satisfies the conditions. If however $O'P\cos O'PQ = OP\cos OPQ$ the admissible value of PR is infinite; in this case we say that the line is *parallel* to the plane.

7·56. It follows that in general there is one point common to three planes.

Lines in a plane are said to be *parallel* if they make the same angle with a given line.

7·57. Parallel lines have no common point at a finite distance. For if (Fig. 10)

$$\angle CAB = \alpha = \angle DBE,$$
$$\angle DBA = \pi - \alpha.$$

If then AC and BD had a common point L, we should have the sum of the angles of the triangle ABL equal to

$$\alpha + (\pi - \alpha) + \angle ALB = \pi + \angle ALB.$$

But this is impossible since A, B, L are not collinear and

$$\angle ALB \neq 0.$$

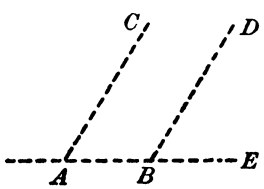

Fig. 10

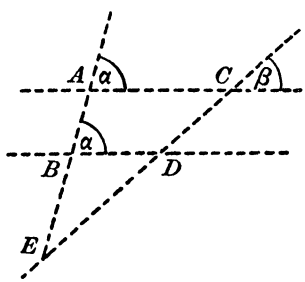

Fig. 11

7·58. Any transversal intersecting the original one makes the same angle with two parallel lines. Suppose the parallels are given to make the same angle α with the line AB, and that CD is another line meeting them (Fig. 11). Let $\angle ACD = \beta$. Then

$$\angle EAC + \angle ECA = \pi - \angle AEC,$$
$$\angle EBD + \angle EDB = \pi - \angle BED.$$

But $$\angle EAC = \angle EBD.$$

Therefore $$\angle ECA = \angle EDB.$$

7·59. If three lines AB, AC, AD are all perpendicular to OA, they are in a plane. For if we make $AO' = OA$, we have

$$BO'^2 = AB^2 + AO'^2 = AB^2 + OA^2 = BO^2,$$

and so on. Hence B, C, and D are all in the plane determined by O and O'.

7·60. Consider any two points L, M and a straight edge OP. Suppose points A on OP, B on OL, C on OM to have been found such that BA, CA are perpendicular to OP. Let

$$OL = r, \qquad OM = r',$$
$$\angle LOP = \theta, \quad \angle MOP = \theta'.$$

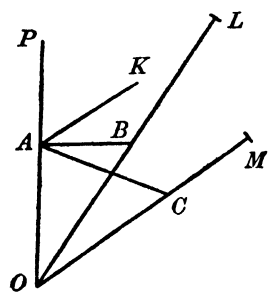

Fig. 12

Then

$$LM^2 = r^2 + r'^2 - 2rr' \cos LOM, \quad (1)$$

$$\cos LOM = \cos\theta \cos\theta' + \sin\theta \sin\theta' \cos BAC, \quad (2)$$

by 7·50. If AK be any other straight edge through A perpendicular to OP, K, A, B, C are in a plane, by 7·59. Let

$$\angle KAB = \phi, \quad \angle KAC = \phi'.$$

Then $$\angle BAC = \phi' - \phi. \quad (3)$$

$$LM^2 = (r\cos\theta - r'\cos\theta')^2 + (r\sin\theta\cos\phi - r'\sin\theta'\cos\phi')^2 + (r\sin\theta\sin\phi - r'\sin\theta'\sin\phi')^2. \quad (4)$$

If now we define x, y, z for L by the equations

$$x = r \sin\theta \cos\phi; \quad y = r \sin\theta \sin\phi; \quad z = r \cos\theta, \qquad (5)$$

we have
$$LM^2 = (x - x')^2 + (y - y')^2 + (z - z')^2. \qquad (6)$$

Thus distance has been expressed in the standard form appropriate to Cartesian co-ordinates.

The angles ϕ and ϕ' are independent of the position of A provided AK is always taken in the same plane. *The angle between two planes* is defined as the angle between two lines in them perpendicular to the common line, and is constant by the corollary to 7·50.

The above definition of Cartesian co-ordinates is applicable in all cases where it is possible to find the distances and bearings of our marks, whereas the usual definition is not applicable unless we can actually find the projections on the three co-ordinate axes. We have still to show that our x, y, z are identical with the usual co-ordinates when these can be measured.

7·61. If we put $x = lr$, $y = mr$, $z = nr$, and consider two marks L, M given by (x_1, y_1, z_1), (x_2, y_2, z_2) we have from 7·60 (2)
$$\cos LOM = l_1 l_2 + m_1 m_2 + n_1 n_2.$$

7·62. At any point of OP, $\theta = 0$, and therefore $l = 0$, $m = 0$, $n = 1$. If OQ is perpendicular to OP, it appears from 7·61 that $n = 0$ at Q. If also Q is in the same plane as OAK, $\phi = 0$ at Q, and $m = 0$. Hence at Q $l = 1$, $m = 0$, $n = 0$. If OR is perpendicular to OP and OQ, then again by 7·61, at R, $l = 0$, $m = 1$, $n = 0$. Thus OQ, OR, OP are the co-ordinate axes as usually understood.

7·63. If $L(x, y, z)$ be another point, the angle between OL and OQ is given by
$$\cos LOQ = l.1 + m.0 + n.0,$$
and therefore by 7·49 the projection of OL on OQ is rl or x. Similar results hold for the projections on the other axes. This gives the identification required. If one or more of l, m, n are

negative, the corresponding foot of the perpendicular from L on the axis lies on the production of the axis beyond O.

7·64. If now a plane is determined by two points (a, b, c), (a', b', c'), we have for all points on the plane

$$(x-a)^2+(y-b)^2+(z-c)^2 = (x-a')^2+(y-b')^2+(z-c')^2; \quad (1)$$

that is,

$$2(a-a')x+2(b-b')y+2(c-c')z = a^2+b^2+c^2-a'^2-b'^2-c'^2. \quad (2)$$

Hence a plane has an equation of the first degree in the co-ordinates. Conversely if we are given an equation of the first degree, which in general involves three independent parameters, we can in a triply infinite number of ways assign the six co-ordinates of O and O' so as to make (2) fit it. Thus every equation of the first degree represents a plane.

It follows that a straight edge is represented by a pair of equations of the first degree. Also if a plane has the equation

$$Ax+By+Cz+D=0, \quad (3)$$

and $P(x_1, y_1, z_1)$, $Q(x_2, y_2, z_2)$ are two points satisfying (3), then

$$R\left(\frac{m_1x_1+m_2x_2}{m_1+m_2}, \frac{m_1y_1+m_2y_2}{m_1+m_2}, \frac{m_1z_1+m_2z_2}{m_1+m_2}\right)$$

also satisfies (3).

If then P and Q are points common to two planes, the point R is also common to the two planes. If we call $R(x, y, z)$ we see that (x, y, z) satisfy

$$\frac{x-x_1}{x_2-x_1}=\frac{y-y_1}{y_2-y_1}=\frac{z-z_1}{z_2-z_1}, \quad (4)$$

the usual form of the equations of a straight line.

From these results the usual analytic development can be carried out.

7·7. The foregoing theory has been developed from the notion and properties of distance alone. Most of the pro-

positions inferred are verifiable by experiment, and have, of course, actually been verified. But other appliances exist that often enable us to supplement the theory and extend its practical application. The first we shall consider is the *protractor*, or graduated circle. A rigid body with one face plane, as tested by the application of a straight edge, is made so as to have a circular edge on this face; that is, every point on the rim is at the same distance from some fixed mark on the face. Equidistant marks are made around the rim, the distance between consecutive marks being compared in the process of manufacture with the length of the turn of a standard screw in much the same way as in the construction of a scale on a straight edge. Then if we consider a triangle formed by the centre and any two consecutive marks on the rim, all such triangles have the same sides, and therefore the same angles. Now if we place the protractor with its centre in contact with the common mark on two intersecting straight edges, and with the rim intersecting both edges, we can count the number of scale-divisions on the rim between the two edges and use it as a measure of the angle between the edges; for either determines the other. Thus the protractor measures angle as a fundamental magnitude.

The actual distance between scale-divisions on a protractor is arbitrary. In practice it is always chosen so that 360 divisions make up the complete circumference and return to the starting-point. If necessary finer graduations are inserted within the original 360. The degree is the angle subtended at the centre by two consecutive divisions on the edge. Now in our measures of angles so far we have specified the angle in terms of its cosine by the series definition of the latter; the number attached to a right angle is $\frac{1}{2}\pi$, and that attached to a complete circumference is 2π. Is an angle, in terms of this measure, merely a number? The test seems to be in attempting addition. 2 sheep and 3 sheep make 5 sheep. But 2 sheep and 3 houses do not make 5 of anything. Now do an angle 2 and the number 3 make 5 of anything? It appears

that they do not. Angle is of a different kind from number, and when we specify it in such a way that the number attached to a right angle is $\frac{1}{2}\pi$ we are really measuring it in terms of a conventional unit, which we call the radian, and is not a number. We should therefore write

$$\tfrac{1}{2}\pi \text{ radians} = 90 \text{ degrees}.$$

This provides the necessary rule for converting measures of angles from one unit to another.

The direct measurement of angle with a protractor now provides a complete substitute for the determination of the angle between straight edges in terms of measured distances. It is still not possible in general when we want the angle XOY and the mark O is not connected to X and Y by actual straight edges. But we can supplement our methods again by using a property of light. It is found that whenever three marks A, B, C are collinear as tested by the straight edge, and the eye is placed so that two of them are in the same direction (a matter of direct sensation) the third is also in that direction. In practice we construct the two nearer marks, for accuracy, as the intersections of crossed threads, so that if the directions do not quite agree small discrepancies will be easily noticeable. Thus we have a direct test of collinearity, which agrees with the test of the straight edge whenever the latter can be applied. We can then generalize this as a test of collinearity and use it instead of the one based on lengths, since it is more accurate and easily applied. Now angles between the directions of marks can be measured. Effectively the crossed wires O in the eyepiece, a distant pair X in the instrument, and the object mark A are placed in a line by the test of coincidence of visual direction; another distant pair Y is placed in line with O and the second object mark B. Then the angle AOB is the same as the angle XOY, which can then be measured with a protractor. The sextant is based on a modification of this principle. The theodolite effectively contains two protractors, and measures two angles corre-

sponding to the θ and ϕ of 7·60; θ is measured from the upward vertical and ϕ from a plane including the upward vertical and the north.

In general three measured data are necessary and sufficient to identify a position. They may be any functions of (r, θ, ϕ) or of (x, y, z) provided that neither of them is a function of the other two. But our principle of simplicity gives reason for regarding the Cartesian co-ordinates as the physically fundamental ones. Our directly recognized entities are straight lines and distances, and a plane is a notion that arises directly out of distance. Now Cartesian co-ordinates have the following simple features possessed by no others. Any plane is expressed by an equation linear in the co-ordinates. Any straight line is expressed by a pair of linear equations; and the co-ordinates of any point on the line are weighted means of those of any two other points on the line. The square of the distance between any two marks is the sum of the squares of the differences between the co-ordinates. No general relations of comparable simplicity hold for any other type of co-ordinates. We regard Cartesian co-ordinates as the physically fundamental ones on account of our principle that the fundamental laws of physics are simple in form.

CHAPTER VIII

NEWTONIAN DYNAMICS

Nature, and Nature's laws, lay hid in night:
God said, Let Newton be! and all was light.
POPE

8·1. Many rigid systems exist. The criterion for a rigid system is that the distances between recognizable marks in it do not change with the time. If one distance in a system and all angles, as tested by optical instruments with graduated circles, do not change with the time, we still call the system rigid by our rules. Over considerable intervals of time most of the objects in this room constitute a rigid system. The angular distances between stars, as observed from the earth, vary with the time so little that decades are required to detect alteration even with the best measuring instruments. If then we consider a system of lines through a given point, and each directed towards a star, that constitutes a rigid system.

When distances change with the time we are in a new realm, called *dynamics*. The marks in one rigid system may change their distances or directions from those in another rigid system. Thus a theodolite and the objects on the earth within its field of view constitute one rigid system; the stars and an equatorial telescope with the clockwork going constitute another; but the directions of the stars change with respect to the theodolite, and those of objects on the earth change with respect to the equatorial.

Objects whose distances and directions with respect to a rigid system are varying with the time are said to have motion relative to the system. Distance and angle have so far been considered only when they are constant for a given set of two or three marks. But even when they vary with the time they still exist. We can specify the position of a particle sliding

down a curve by the mark on the curve that the particle is passing over. We can specify the direction of a planet by pointing an equatorial telescope towards it, and reading its right ascension and declination on the graduated circles; or we can photograph the region of the sky where it is, and measure its angular distances from neighbouring fixed stars just as we can measure the angular distances between these stars themselves. In such a case as the ascent of a pilot balloon, observed with two theodolites, we can actually observe the directions from two positions simultaneously and determine the position of the balloon at each instant of observation just as for a fixed object. In dynamics we are therefore dealing with cases where distances, and those entities we have found to depend on them, still exist, but are now functions of the time instead of being constant.

One clearly cut distinction arises immediately. In most rigid systems there is a continuous material connexion, traceable by sight and touch, between all parts. If no such connexion is evident, as in a body in mid-air, or a planet, or the components of a double star, there is in general motion relative to other rigid systems. There is therefore a strong suggestion that material connexion between bodies is antagonistic to relative motion. Even if there is relative motion to begin with, as in the case of a body projected along the floor, it soon stops when material contact is established. We shall not at present examine the nature of this phenomenon; we merely give it a name. The property that one body does not move through another we call *impenetrability*; the property that relative motion tends to cease when one body slides over another we call *friction*. In dynamics, then, we refer our measurements to some rigid system, in which the laws relating measurements, whether made at the same or at different times, are already known; but our subject-matter is the motion with reference to our rigid system of a body or bodies that are not constrained by material connexion to have no motion with reference to it. An immediate inference is that we should

consider in the first place those bodies that have the slightest possible material connexion with our frame of reference; in this way we reduce to a minimum the interference of material connexion with motion and remove one independent variable.

8·2. This condition is obviously satisfied fairly well by bodies in mid-air, and very well by the heavenly bodies. In the former case we take as our frame of reference axes fixed in the earth, and find that the motion, at any rate for massive bodies, is well represented by the differential equations

$$\frac{d^2x}{dt^2} = 0; \quad \frac{d^2y}{dt^2} = 0; \quad \frac{d^2z}{dt^2} = -g, \tag{1}$$

where x and y are measured horizontally and z vertically upwards, as judged by a plumb-line. The derived magnitude g is nearly constant. Now by simultaneous observations of the stars from different places on the earth's surface we find that the direction of the plumb-line is not everywhere the same with reference to the directions of the stars, but points nearly to a fixed point in the earth, which we call the centre. If we take new Cartesian co-ordinates with respect to the centre of the earth as origin we now find that, wherever we are on the earth's surface, the equations (1) lead to

$$\frac{d^2x}{dt^2} = -g\frac{x}{r}; \quad \frac{d^2y}{dt^2} = -g\frac{y}{r}; \quad \frac{d^2z}{dt^2} = -g\frac{z}{r}, \tag{2}$$

where r is measured from the centre of the earth. This is a more general form than (1). It leads to a further suggestion, that the second derivatives of the Cartesian co-ordinates with respect to the time are of fundamental importance in dynamics; for they are expressed by three known functions of the co-ordinates themselves, and lead thereby to three differential equations for these co-ordinates. We call the first derivatives of the Cartesian co-ordinates the components of relative velocity, and the second derivatives the components of relative acceleration.

8·21. Now consider the motion of the components of a double star. We take an axis of x in a direction fixed with reference to the directions of the majority of the stars, and such that the double star is always near it. The axes of y and z are taken in two directions perpendicular to each other and to that of x. We observe the angles between the direction of a component of the star and the two planes of xy and xz; or, what is equivalent, we take the point of intersection, with a plane plate perpendicular to the x axis, of the line joining the centre of the object glass of the telescope to the component. It is found that as time goes on the points given by the two components describe similar ellipses. If we put $y/x = p$, $z/x = q$, and use suffixes 1 and 2 for the two components, the variations of p and q in the same interval of time are always opposite in direction and always in the same ratio, except for a uniform velocity shared by both components of the star. If we proceed to the second derivatives to remove this uniform part of the rate of change, we have

$$\frac{\ddot{p}_1}{\ddot{q}_1} = \frac{\ddot{q}_2}{\ddot{p}_2}. \tag{1}$$

Further, each ratio is equal to $\dfrac{p_2 - p_1}{q_2 - q_1}$. Now p and q are always small, and the displacements at right angles to the line of sight are therefore small fractions of the whole distance. We must choose between two alternatives with regard to the displacements in the line of sight. If they are also small compared with the distance of the star, the variation of x is small compared with its mean value, and y and z for each component are nearly proportional to p and q. Then

$$\frac{\ddot{y}_1}{\ddot{z}_1} = \frac{\ddot{y}_2}{\ddot{z}_2} = \frac{y_2 - y_1}{z_2 - z_1}. \tag{2}$$

The alternative is that the displacements in the line of sight are comparable with the distance; this would mean that the orbit of every double star is enormously elongated towards the earth, and we need not consider this possibility seriously.

Returning to (2) now, we see that there is probably nothing special about the line of sight and we can generalize the equations in the form

$$\frac{\ddot{x}_1}{x_2-x_1}=\frac{\ddot{y}_1}{y_2-y_1}=\frac{\ddot{z}_1}{z_2-z_1};\quad \frac{\ddot{x}_2}{x_2-x_1}=\frac{\ddot{y}_2}{y_2-y_1}=\frac{\ddot{z}_2}{z_2-z_1}. \quad (3)$$

The components of acceleration are in the ratios of the differences of the co-ordinates; in other words, the accelerations of the bodies are along the line joining them. Further, we can choose a ratio of two quantities m_1 and m_2 such that

$$m_1\ddot{x}_1+m_2\ddot{x}_2=0;\ m_1\ddot{y}_1+m_2\ddot{y}_2=0;\ m_1\ddot{z}_1+m_2\ddot{z}_2=0. \quad (4)$$

Distant bodies appear to produce no acceleration on one another, as is seen from the negligible or constant velocities of most of the stars. Hence we can say that the acceleration of each component is *due to* the proximity of the other component.

Similar results are found for most of the satellites of the planets; they are consistent with the acceleration in each case being directed towards the centres of the planets.

8·22. Now consider the acceleration of the moon. To a first approximation the moon describes a circle about the earth with radius a and angular velocity n. The acceleration in such a path is an^2 towards the centre of the earth. Taking

$$a=3{\cdot}8\times10^{10}\text{ cm.},\quad n=2\pi/27{\cdot}3\text{ days},$$

we find that the acceleration is 0·273 cm./sec.^{2} Now a particle at the earth's surface has acceleration 980 cm./sec.^{2}, which is nearly 3600 times the acceleration of the moon. The distances are in the ratio 1 : 60 nearly, and therefore the accelerations are nearly inversely as the squares of the distances.

8·23. These few facts relating to freely moving bodies suggest the following summary:

A body has an acceleration in the direction of a neighbouring body, and proportional in magnitude to the inverse square of the distance.

The accelerations that two bodies produce on each other are in a ratio independent of the time.

The first of these laws can be written in the form

$$\ddot{x}_1 = -\frac{\mu_2}{r^2}\frac{x_1 - x_2}{r}; \quad \ddot{y}_1 = -\frac{\mu_2}{r^2}\frac{y_1 - y_2}{r}; \quad \ddot{z}_1 = -\frac{\mu_2}{r^2}\frac{z_1 - z_2}{r}, \quad (1)$$

where (x_1, y_1, z_1) are the co-ordinates of the body whose acceleration we want, (x_2, y_2, z_2) those of the other body, and μ_2 is a constant of proportionality independent of the co-ordinates and of t. The second law then implies that

$$\ddot{x}_2 = \frac{\mu_1}{r^2}\frac{x_1 - x_2}{r}; \quad \ddot{y}_2 = \frac{\mu_1}{r^2}\frac{y_1 - y_2}{r}; \quad \ddot{z}_2 = \frac{\mu_1}{r^2}\frac{z_1 - z_2}{r}, \quad (2)$$

where μ_1 is a second constant of proportionality, different in general from μ_2.

This family of differential equations can be solved exactly. They are found to imply the following consequences:

A point with co-ordinates $(\bar{x}, \bar{y}, \bar{z})$ given by

$$(\mu_1 + \mu_2)\,\bar{x} = \mu_1 x_1 + \mu_2 x_2, \quad (3)$$

and two similar equations, moves with uniform velocity in a straight line. We call this point the centroid of the two particles.

Relative to this point both the bodies describe ellipses, the ellipses being similar but having their axes in opposite directions, and the centroid being in a focus of each.

The line joining the centroid to either body sweeps out in any interval of time an area proportional to that interval.

The mean distance a between the bodies being defined as the mean of their greatest and least distances apart, and the mean motion n as 2π divided by the time of describing the orbit,

$$n^2 a^3 = \mu_1 + \mu_2. \quad (4)$$

These results express Newton's solution of the Problem of Two Bodies. It is found to describe accurately the motions of double stars. Only motions at right angles to the line of

sight being measurable*, what we actually verify is that the movements agree with the projections of elliptic motions that follow the laws. In other words, the behaviour of two of the three variables determined by the solution is completely verified; our present analysis is neither confirmed nor contradicted by observations of the other variable, for these do not exist.

When we consider the motions of satellites about their primaries, the same solution is found to fit the relative motion, as to the two measurable co-ordinates. Also the planet itself shows no departure from a regular motion relative to the stars; over intervals of time amounting to several periods of revolution of the satellites the motion of the planet is sensibly uniform. It appears therefore that the centroid of the planet and any satellite is practically coincident with the centre of the planet, and therefore that if μ_1 refers to the planet and μ_2 to the satellite, μ_2/μ_1 is always very small and $\mu_1 + \mu_2$ is practically μ_1. But for different satellites of the same planet we get a further check; the quantity n^2a^3 is found to be the same for all. The constancy of n^2a^3 for different satellites also therefore implies that μ_1 is a property of the planet.

Coming now to the motions of the sun and planets, we can observe in each case only directions as seen from the earth with reference to the stars, except in the case of the sun, where we can estimate the variation of its distance by measuring its angular diameter from time to time. In this case then we can check all three co-ordinates, and we find that the motion of the sun relative to the earth is definitely an ellipse with the earth in a focus. For the other planets it is found that ellipses can always be found with the sun in a focus, such that the radius vector relative to the sun sweeps out area at a uniform rate, and the direction of the planet as seen from the earth agrees with that predicted from the various elliptic

* That is to say, in terms of the considerations of direction that we have used so far. Velocities in the line of sight can be measured by means of the Doppler effect, and agree with the laws, but we are not yet in a position to discuss the theory of that effect.

paths. It may be noticed that each elliptic orbit is specified when six quantities are given, namely the three co-ordinates and the three components of velocity relative to the sun at some definite instant. With only these six adjustable parameters it is possible to fit an indefinitely large number of observations of direction as the planet describes its orbit. The alternative to supposing that the motion is actually a Newtonian elliptic orbit is that the distance of the planet from the earth does not follow the rules found for elliptic motion, but that the co-ordinates are so related that the direction does satisfy these rules. As there is no intelligible reason why this should be true apart from the truth of the equations of motion, we do not treat this alternative seriously.

It is now found that for each planet the quantity n^2a^3 is the same. As for satellites, we therefore argue that it expresses a property of the sun, and that μ for each planet is very small compared with its value for the sun. This can be checked directly, since the values of μ for those planets that have satellites are already known from observations of the motions of the satellites relative to their primaries. Also it is found that the value of n^2a^3 found for the motion of the sun relative to the earth is the same as that found for the motions of the other planets relative to the sun. It therefore expresses a property of the sun rather than the earth, and we say that all the planets, the earth included, describe elliptic orbits about the sun. This is legitimate because the co-ordinates of the earth relative to the sun, the directions of the axes remaining the same, are necessarily equal and opposite to those of the sun relative to the earth, so that if the sun describes relative to the earth an elliptic orbit with the earth in a focus, then the earth also describes, relative to the sun, an elliptic orbit with the sun in a focus*.

* Copernicus and Kepler laboured under the disadvantage of having no accurate observations of double stars or satellites among their data. Jupiter's greater satellites were discovered in 1609, the year of the publication of Kepler's first two laws, but their orbits are nearly circular. The same applies to the two largest of those of Saturn. Accurate demonstration

8·24. We have seen therefore that a body in the neighbourhood of a second body has an acceleration towards the second body, inversely proportional to the square of the distance apart, the constant of proportionality being a property of the second body alone. What happens when there are several bodies in the neighbourhood? We really have had the answer already in the motions of satellites; for while a satellite is moving about its primary it is also sharing the general motion of the primary about the sun. If its acceleration was merely that towards the primary, while the primary is moving about the sun, the primary would leave the satellite behind*. The satellite must have also an acceleration towards the sun, which is nearly the same as that of the primary because they are at nearly the same distance from the sun. We must therefore generalize our law to the case where n bodies are moving in one another's neighbourhood. We say that any one body has an acceleration towards each of the others, whose components are given by our law; and the total component acceleration in any direction is the sum of the components in that direction given by the other bodies separately. Formally we say that if we consider the lth body,

$$\ddot{x}_l = -\Sigma \frac{\mu_m (x_l - x_m)}{r^3_{lm}}; \quad \ddot{y}_l = -\Sigma \frac{\mu_m (y_l - y_m)}{r^3_{lm}};$$

$$\ddot{z}_l = \Sigma \frac{\mu_m (z_l - z_m)}{r^3_{lm}}, \quad (5)$$

where the suffix m refers to another body of the system, r_{lm} is the distance between the bodies specified by l and m, and the summation is for all values of m except l.

of the elliptic motion in double stars is a matter of comparatively modern observation. If Kepler had had such data, it would not have taken him six years to hit on the elliptic law of planetary motions. As it was, he had to tackle directly the more difficult problem of the motions of the planets, in which the complicating influence of the earth's motion is seen at its worst.

* This is serious; the acceleration of the moon towards the sun, for instance, is about twice its acceleration towards the earth.

We notice that

$$r^2_{lm} = (x_l - x_m)^2 + (y_l - y_m)^2 + (z_l - z_m)^2, \quad (6)$$

and

$$\frac{x_l - x_m}{r^3_{lm}} = -\frac{\partial}{\partial x_l}\frac{1}{r_{lm}} = \frac{\partial}{\partial x_m}\frac{1}{r_{lm}}, \quad (7)$$

with similar relations. Then if we multiply the equations (5) respectively by $\dot{x}_l, \dot{y}_l, \dot{z}_l$ and add, we get

$$\dot{x}_l\ddot{x}_l + \dot{y}_l\ddot{y}_l + \dot{z}_l\ddot{z}_l = \Sigma \dot{x}_l \frac{\partial}{\partial x_l}\frac{\mu_m}{r_{lm}} + \Sigma \dot{y}_l \frac{\partial}{\partial y_l}\frac{\mu_m}{r_{lm}} + \Sigma \dot{z}_l \frac{\partial}{\partial z_l}\frac{\mu_m}{r_{lm}}. \quad (8)$$

The left side is a complete differential with regard to the time. If then we integrate from time t_0 to time t_1 we get

$$\left[\tfrac{1}{2}(\dot{x}_l^2 + \dot{y}_l^2 + \dot{z}_l^2)\right]_{t_0}^{t_1} = \int_{t_0}^{t_1}\left(\frac{\partial U_l}{\partial x_l}dx_l + \frac{\partial U_l}{\partial y_l}dy_l + \frac{\partial U_l}{\partial z_l}dz_l\right), \quad (9)$$

where

$$U_l = \Sigma \frac{\mu_m}{r_{lm}}. \quad (10)$$

We can also multiply (8) by μ_l and add for all values of l. Then the pair of particles given by l and m make a contribution to the right given by

$$\mu_l\mu_m\left(\dot{x}_l\frac{\partial}{\partial x_l} + \dot{x}_m\frac{\partial}{\partial x_m} + \dot{y}_l\frac{\partial}{\partial y_l} + \dot{y}_m\frac{\partial}{\partial y_m} + \dot{z}_l\frac{\partial}{\partial z_l} + \dot{z}_m\frac{\partial}{\partial z_m}\right)\frac{1}{r_{lm}} = \mu_l\mu_m\frac{d}{dt}\frac{1}{r_{lm}}, \quad (11)$$

since r_{lm} is a function of $(x_l, y_l, z_l, x_m, y_m, z_m)$ only. It follows that if

$$U = \Sigma \frac{\mu_l\mu_m}{r_{lm}}, \quad (12)$$

where the summation is for all pairs of particles,

$$\left[\Sigma\, \tfrac{1}{2}\mu_l(\dot{x}_l^2 + \dot{y}_l^2 + \dot{z}_l^2)\right]_{t_0}^{t_1} = \left[U\right]_{t_0}^{t_1}. \quad (13)$$

This is a very remarkable result. For the expression

$$\tfrac{1}{2}(\dot{x}_l^2 + \dot{y}_l^2 + \dot{z}_l^2)$$

is the square of the resultant velocity of the particle given by l; where we consider the positions of the particle at times t and $t + dt$, and take the distance between them, and define the resultant velocity as the limit of the ratio of this distance to dt when dt becomes very small. The quantities μ are properties of the various bodies and independent of their position. Thus the equation (13) expresses a relation between our fundamental notions of distance and time alone, and is independent of the particular set of axes of x, y, and z that we choose. Further, (5) are equivalent to

$$\mu_l \ddot{x}_l = \frac{\partial U}{\partial x_l}; \quad \mu_l \ddot{y}_l = \frac{\partial U}{\partial y_l}; \quad \mu_l \ddot{z}_l = \frac{\partial U}{\partial z_l}. \qquad (14)$$

The generalization (5) makes a great improvement in our representation of the motions within the solar system. Kepler's laws give a good first approximation to the motions of the planets; their application to the motions of the satellites relative to the planets also gives a good first approximation. But there are outstanding discrepancies. There are periodic inequalities in the moon's longitude with amplitudes of the order of a degree; others in the longitudes of the planets of the order of, in extreme cases, considerable fractions of a degree; there is a long-period disturbance of Saturn with a period of 900 years and an amplitude of nearly a degree; and in addition the elements of the orbits show slow progressive or secular changes, the major axes in particular revolving in one direction or the other relative to the stars. The result of allowing for the mutual influence of *every* pair of bodies in the system is that nearly all these inequalities are accounted for. Without further modification we can account, within the limits of observational error, for the motion of every major planet from Venus to Neptune, all the asteroids, and most of the satellites.

The outstanding discrepancies all concern cases where the body whose motion we are considering is very near its primary. This fact suggests an explanation; for we have seen

that the acceleration of a body is always towards the body that produces it. If the latter is remote, the lines joining the first body to all parts of the second are nearly in the same direction. But if the bodies are fairly close together these lines are not in the same direction, and if the second body is not spherical its field will not be symmetrical, and the acceleration of the first body will not necessarily be directed towards a fixed point of the second. Such considerations do as a matter of fact account for most of the outstanding inequalities of the satellites. The only remaining inequality of importance concerns Mercury; its discussion is reserved till later.

We have seen that the acceleration of any body can be considered as made up of contributions from the others, each of which can be said to be *due to* another particular body in the sense that it would be zero if the other body were not present. If then we denote the part of the acceleration of the particle l due to the particle m by $\ddot{x}_{lm}, \ddot{y}_{lm}, \ddot{z}_{lm}$ we have

$$\mu_l \ddot{x}_{lm} + \mu_m \ddot{x}_{ml} = 0, \text{ etc.} \qquad (15)$$

We can call the terms in this equation the respective forces of the bodies on each other, and we arrive at a result equivalent to Newton's third law, that action and reaction are equal and opposite.

The equation (15) is probably most directly verified in the solar system by the mutual perturbations of Jupiter and Saturn, and by those of the four great satellites of Jupiter. The disturbance of the position of the sun by the attractions of the planets affects the positions of the planets relative to it, but cannot be disentangled explicitly. The earth and moon move about their common centre of gravity, which moves practically like a single particle. There is therefore a monthly oscillation of the earth's position, which is shown by a corresponding variation in the apparent direction of the sun, and gives a means of determining μ for the moon. But there is no other way of finding μ for the moon from the translational motions of bodies, so that this determination is at present

merely an application of the principle and not a check on it.

We may remark that there is nothing conventional about the quantities μ. They are perfectly definite derived magnitudes. Thus μ for the sun $= 4\pi^2 \dfrac{(\text{earth's mean distance})^3}{(1 \text{ year})^2}$. The notion of mass has not yet appeared explicitly.

8·3. There is one apparent inconsistency in the development given so far. In considering the motion of a body near the earth's surface, we referred it to an origin at the centre of the earth and axes fixed in the earth. In considering the motions of the bodies in the solar system we have used axes whose directions are fixed with reference to the stars. But axes fixed in the earth do not keep the same direction with reference to the stars, or conversely; and it is easy to see that the equations of motion cannot keep the same form if the axes are rotating. Thus our equations

$$(\ddot{x}, \ddot{y}, \ddot{z}) = -\frac{\mu}{r^3}(x, y, z) \tag{1}$$

are satisfied as they stand if

$$x = a\cos\omega t; \quad y = a\sin\omega t; \quad z = 0, \tag{2}$$

where a and ω are constants such that

$$\omega^2 a^3 = \mu. \tag{3}$$

But if we take axes of (x', y', z') rotating about the z axis with angular velocity ω, the co-ordinates in the two systems are connected by relations

$$x' = x\cos\omega t + y\sin\omega t; \quad y' = -x\sin\omega t + y\cos\omega t; \quad z' = z. \tag{4}$$

Then

$$x' = a, \quad y' = 0, \quad z' = 0, \tag{5}$$

but in these co-ordinates the equations of motion in the form (1) are not satisfied, for the first reduces to the impossible form

$$0 = -\omega^2 a. \tag{6}$$

It appears therefore that we cannot retain the same form of the equations of motion for all sets of axes in relative rotation. If the form (1) is true for axes with directions fixed in relation to the stars, it cannot be correct for axes fixed in relation to the earth, and conversely. Now our study of the motions within the solar system has been made with reference to axes fixed with reference to the stars. If we assume this to be true in general a modification is needed for axes fixed in the earth, which is given in books on dynamics. It is actually found to be small for a projectile moving near the earth's surface, really because the time of flight is always so small that the earth rotates through only a small angle during it. But the correction is appreciable in long-range gunfire, and has to be taken into account in accurate shooting. The equations therefore hold for axes fixed in direction in relation to the stars.

In the last resort this statement requires to be made a little more precise; for, though the angles between the directions of the stars are nearly constant, they are not quite so. The stars have slow proper motions among themselves, and if we fix the directions of our axes with regard to one pair of stars they will vary with regard to another pair. In practice however it is found that there are an abundance of stars the angles between whose directions remain constant as nearly as we can observe. These are the distant stars, and we use these as our general standards of direction. But strictly, even the most distant stars must have accelerations on account of the law itself, and we can never identify absolutely non-rotating axes. This does not affect our belief in the truth of the law, of course. The law is approximately true as a matter of observation, and fits the observations as closely as we can tell; given these properties, the law has a high probability because it is simple.

The position of the origin has been left somewhat vague. It is plain that if we take a different origin moving with a uniform velocity with respect to the first one, the co-ordinates are merely reduced by quantities of the form $a + ut$, which

are the same for all particles. Hence quantities of the forms $\ddot{x}$ and $x_l - x_m$ are exactly as before. The equations of motion are unaffected by a displacement or a uniform velocity of the origin. This is Newton's *principle of relativity*. But actually we may take a new origin moving in any way whatever. If we do so, the quantities $x_l - x_m$, and therefore all distances, remain unaltered; x may be changed, but by the same amount for all bodies, and therefore $x_l - x_m$ is unchanged. The differential equations for the differences of the co-ordinates of the various bodies remain as before. But actually we can never observe actual co-ordinates; all we observe are the differences of the co-ordinates. It appears therefore, that as far as actual tests are concerned the origin may move in any way whatever and the equations of motion will still lead to correct results for the quantities that we can observe.

There may, however, be advantages in having equations of motion that are actually true, rather than such as contain errors that can never be discovered. If we consider the point with co-ordinates $(\bar{x}, \bar{y}, \bar{z})$, which we may call the centroid of the universe, defined by

$$(\Sigma\mu_l)\,\bar{x} = \Sigma\mu_l x_l, \tag{1}$$

with similar equations, we have

$$(\Sigma\mu_l)\,\ddot{\bar{x}} = \Sigma\mu_l\ddot{x}_l = \sum_{l,m}(\mu_l\ddot{x}_{lm} + \mu_m\ddot{x}_{ml}) = 0, \tag{2}$$

if the equations of motion are true as they stand. Then the centroid moves with uniform velocity with reference to the origin. Conversely, if our origin moves with uniform velocity with reference to the centroid of the universe, the equations of motion are true. We notice that if we shift the origin to the centroid, the new co-ordinates take the form

$$x_l' = x_l - \bar{x} = \frac{(\Sigma\mu_l)\,x_l - \Sigma\mu_m x_m}{\Sigma\mu_l} = \frac{\Sigma\mu_m\,(x_l - x_m)}{\Sigma\mu_m}. \tag{3}$$

Thus the equations in this form involve differences of the co-ordinates alone.

With these co-ordinates

$$\Sigma\mu_l\dot{x}_l^2 = \Sigma\mu_l\dot{\bar{x}}^2 + 2\dot{\bar{x}}\,\Sigma\mu_l\dot{x}_l' + \Sigma\mu_l\dot{x}_l'^2, \qquad (4)$$

and the second term vanishes by (1). Also $\dot{\bar{x}}$ is constant or zero. Thus

$$\Sigma\mu_l(\dot{x}_l^2 + \dot{y}_l^2 + \dot{z}_l^2) = \Sigma\mu_l(\dot{\bar{x}}^2 + \dot{\bar{y}}^2 + \dot{\bar{z}}^2) + \Sigma\mu_l(\dot{x}_l'^2 + \dot{y}_l'^2 + \dot{z}_l'^2), \qquad (5)$$

in which the first term is constant if we have chosen the origin suitably.

8·4. When we come to deal with bodies that are not moving freely we find a difference at once. The acceleration becomes infinite when the bodies become indefinitely close if the foregoing equations are true. Actually when two bodies come into contact the relative acceleration disappears; the law of attraction undergoes serious modification at this stage*. What are we to say about a book lying on a table? Two courses are available. The book is not in contact with the earth; we may say that due to the earth the book has, as usual, the acceleration g downwards, but owing to the proximity of the table it has also an acceleration g upwards, and the two cancel. Otherwise we may just say that the acceleration is zero and leave it at that. It is in many cases a matter of convenience which course we adopt; both give the same answer as far as observable phenomena are concerned. But the discussion of the additional reactions associated with impenetrability and friction has the important feature that it brings out new physical laws.

Consider a common balance with a fixed counterpoise in the pan A. We place various bodies in the pan B. According

* Humpty-Dumpty would, of course, have been quite wrong had he said that *Impenetrability* meant a nice knock-down argument. The impenetrability of the wall was what was keeping him in position in spite of gravity; it was when a gust of wind removed him from its range of influence that he became a freely moving body and could be said to be knocked down.

to our results for freely moving bodies, each of these has its appropriate μ, and each has an acceleration g downwards due to the earth. Then in terms of equation (1) there is something associated with the effect of the earth on each body that is expressed by the quantity μg. If the balance is held with the counterpoise in place and released, the counterpoise goes down and the pan B rises. If the experiment is repeated with various bodies in the pan B, some rise and others fall when the balance is released. It appears that the effect of the balance on some bodies is enough to overcome μg, and in others is not. But in each case the balance itself, with the counterpoise, starts off from the same conditions. Thus the operation of using the balance classifies bodies according to their values of μg or, since g is the same for all, according to their values of μ.

We have a check on this. If we return to the problem of the solar system and suppose that the distances between the bodies specified by $m_1, m_2, m_3, \ldots m_n$ are all small compared with their distance from the body specified by l, the acceleration of the last due to all together is nearly

$$(\mu_{m_1} + \mu_{m_2} + \ldots + \mu_{m_n})/r^2_{lm}$$

towards the centroid of all the particles. Thus for particles close together the effects on another body are expressible by treating μ as an additive quantity. If we place several bodies in the pan of a balance at once, similarly, the effect of the balance on all together is in opposition to the effects of the earth on all together; and since we must measure the effect of all on the earth by the sum of the values of μg, we naturally measure the effect of the earth on them also by the sum of the values of μg.

It appears therefore that a balance with a standard counterpoise provides a means of discriminating between bodies, or combinations of bodies, according to the sums of their respective μ's; and these sums have the additive property. It follows that the mass m of a body, determined by the balance

as a fundamental magnitude, is proportional to the derived magnitude μ. Being a fundamental magnitude, mass has to be measured in terms of a unit; the value of μ for this unit mass remains to be determined. We shall denote it by f, so that $\mu = fm$. We can now generalize the notion of mass to bodies too large or too inaccessible to be weighed on a balance, by saying that it is proportional to μ, which exists in general. Then our law 8·24 (15) relating the effects of two bodies on each other takes the form, for any pair of masses m_1 and m_2,

$$m_1\ddot{x}_{12} + m_2\ddot{x}_{21} = 0.$$

In this form we call $m_1\ddot{x}_{12}$ the *force* on m_1 due to m_2, and $m_2\ddot{x}_{21}$ the force on m_2 due to m_1. Then we have Newton's third law in its usual form, that the forces on two bodies due to each other are equal and opposite. We may denote them respectively by X_{12} and X_{21}.

We have also the law that the acceleration of a body is the sum of those due to the other bodies in the world, obtained by adding the components in Cartesian co-ordinates. If we simply multiply this equation by the mass of the body, we have the rule for the composition of forces due to different bodies, that the total force on a body is the resultant of the forces due to the other bodies, obtained by adding their components in Cartesian co-ordinates.

The reasons why force is interesting are then, first, that it has the symmetrical property that action and reaction are equal and opposite; second, that the forces acting on a body due to other bodies are additive; third, that when the state of a system is known the forces are found to be determinate functions of the co-ordinates and possibly the velocities. Thus the equations of the form $m\ddot{x} = X$ are strictly differential equations for the co-ordinates.

Strictly these results have been established only for freely moving bodies. We proceed to extend them to bodies in general, even when the phenomena of impenetrability and friction arise. Their complete verification is then impossible,

because motion in one co-ordinate is always prevented by the connexion of the apparatus with the earth, and though there may be a force between the apparatus and the earth we can never measure the corresponding acceleration of the earth.

But their partial verification is easy. Consider a body of mass m_1 hanging by a string of mass m_2 from the pan of a balance, the whole being at rest. The body is acted on by a force $m_1 g$ downwards, which we call its weight. It has no acceleration. Therefore the string is producing on it a force $m_1 g$ upwards. If action and reaction are equal and opposite, the body is therefore producing a downward force $m_1 g$ on the string. Also the weight of the string is $m_2 g$, so that there is a total downward force $(m_1 + m_2)\, g$ acting on the string. But the string has no acceleration and must therefore be acted on by an upward force $(m_1 + m_2)\, g$ from the balance. Therefore the pan of the balance is subject to a downward force $(m_1 + m_2)\, g$. But if we untie the string and place the string and body in the pan of the balance the counterpoise is undisturbed. The force on the balance pan in the two cases is therefore the same; and in the second case we know directly that it is equal to the sum of the weights of the two bodies. We have therefore a verification of a direct inference from the laws.

8·5. Now consider a body with a plane face resting on an inclined plane at an inclination α to the horizontal. It is subject to a vertical acceleration g due to the earth. This is equivalent, whatever axes are chosen, to an acceleration $g \cos \alpha$ normally into the plane and another $g \sin \alpha$ down the plane. If the body remains at rest, these must be balanced by reactions due to the plane. If m is the mass of the body, there must therefore be acting on it a force $mg \cos \alpha$ normally and one $mg \sin \alpha$ up the plane. We call these the normal and frictional reactions. The first preserves impenetrability, the second prevents slip.

But if the slope of the plane is gradually increased it is found that at a certain inclination λ the body can no longer remain at rest, but slides down the plane. At greater inclina-

tions it has an acceleration $g(\sin\alpha - \tan\lambda\cos\alpha)$. The first term corresponds to the component of gravity. The second shows that there is a frictional reaction $mg\tan\lambda\cos\alpha$, acting against the motion. There is still no normal acceleration, and therefore the normal reaction is still $mg\cos\alpha$. The ratio of the frictional reaction to the normal one is therefore $\tan\lambda$ when $\alpha > \lambda$ and $\tan\alpha$ when $\alpha < \lambda$. The constant λ is a property of the nature of the surfaces in contact.

But if the body is projected up the plane, or if it is projected down the plane when $\alpha < \lambda$, it is found that the portions $mg\sin\alpha$ and $mg\tan\lambda\cos\alpha$ behave differently. The former always acts down the plane. The latter acts against the direction of motion, whatever that may be. If we take the body by hand and slide it about the plane by applying a force parallel to the plane, the former force is helping us when we push the body down the plane, and opposing us when we push it upwards. The latter is always opposing us. This introduces us to the distinction between conservative and non-conservative forces, which requires further elucidation.

The equation 8·24 (13), with a suitable origin, is equivalent to

$$\left[\Sigma\tfrac{1}{2}m_l\,(\dot{x}_l^2 + \dot{y}_l^2 + \dot{z}_l^2)\right]_{t_0}^{t_1} = \left[\Sigma f\frac{m_l m_m}{r_{lm}}\right]_{t_0}^{t_1}. \qquad (1)$$

In this form we may call the left side the *kinetic energy* of the system, and the contribution to it from each body the kinetic energy of that body. If we call the expression on the left T, and that on the right U, then however the system may move $T - U$ remains constant. If it should happen that the system ever gets back to its initial position, it will have its initial kinetic energy again.

If the total force acting on any body is (X, Y, Z), and we start from the three equations of motion of the type

$$m\ddot{x} = X, \qquad (2)$$

we can infer

$$\left[\tfrac{1}{2}m\,(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)\right]_{t_0}^{t_1} = \int_{t_0}^{t_1}(X\dot{x} + Y\dot{y} + Z\dot{z})\,dt. \qquad (3)$$

We express this in words by saying that the increase in kinetic energy is equal to the *work done* on the body. If we have a system of bodies we can add the equations of this type for all, and say that the increase of kinetic energy of the system is equal to the total work done on all the bodies. In any case this applies to the actual path. But we have seen that the forces (X, Y, Z) are determinable as functions of the positions and velocities. However we imagine the system to travel from its initial position to its final one, by whatever path and at whatever rate, the integral on the right of (3) has some value, provided we give (X, Y, Z) the proper values for a body with the co-ordinates and components of velocity that we are considering. It happens in many cases that the value is the same whatever path we choose, provided the initial and final positions are the same. This is true, for instance, in the motion of the bodies of the solar system. If so, the work done is a function of the initial and final positions only and not of the path taken. Such a system is called *conservative*.

In the case of the body on the inclined plane, if we imagine it displaced a distance s down the plane, the work done is $mgs\,(\sin\alpha - \tan\lambda\cos\alpha)$. If it is then brought back to the starting point, the force is now $mg\,(\sin\alpha + \tan\lambda\cos\alpha)$ *against* the direction of motion, and the work done is $-mgs\,(\sin\alpha + \tan\lambda\cos\alpha)$. Adding the two together we have the total work done, $-2mgs\tan\lambda\cos\alpha$. This depends on s and therefore on where the body has been in passing from its initial to its final position. The system is not conservative.

It appears that the work done by non-conservative forces is always negative. It is found also that, associated with it, there is a change in the state of the system, which we call heating, and can detect by direct sensation or by a thermometer.

8·6. A further refinement must be introduced at this stage. We have proceeded so far by supposing that the position of a real body can be expressed by three co-ordinates (x, y, z).

This is not actually true, because a body has finite size and may have rotation. But in fact co-ordinates are obtained in the process of measurement, and we must examine the meaning of the co-ordinates we have obtained. For a planet or a star we have not watched a particular mark on the surface; our result is really that there *is* a point within the body whose co-ordinates do satisfy our equations. Actual marks on the surface have additional accelerations in consequence of the rotation, and do not satisfy the equations as they stand*. In reality such expressions as $m\ddot{x}$ have no meaning unless the value of $\ddot{x}$ can be treated as uniform throughout the region considered. But then the region is in general only a small portion of that occupied by the body, and we should take into account the forces due to the other portions of the body that surround it.

The principle actually used is that the internal reactions between portions of a body constitute a system in equilibrium. There seem to be several grounds given for accepting this. If we consider any particle at (x, y, z), we can write its equations of motion in the form

$$m\ddot{x} = X + X', \tag{1}$$

where X is the force on it due to external bodies and X' that due to other parts of the same body. If we add up these equations for all particles we get

$$\Sigma m\ddot{x} = \Sigma X + \Sigma X'. \tag{2}$$

Also from the equations of the form (1) we can obtain three of the form

$$\Sigma m(y\ddot{z} - z\ddot{y}) = \Sigma(yZ - zY) + \Sigma(yZ' - zY'). \tag{3}$$

Now if we consider any pair of particles m_1 and m_2, their forces on each other are equal and opposite. Hence in the

* A particle at the earth's equator has an acceleration of 3·4 cm./sec.^{2} towards the axis on account of rotation; the general acceleration of the earth towards the moon is $3{\cdot}4 \times 10^{-3}$ cm./sec.^{2}.

sum $\Sigma X'$ the forces cancel in pairs, and the total is zero. Also since the forces do act on both particles, they must act along the line joining them*. Hence if the particles are at (x_1, y_1, z_1) and (x_2, y_2, z_2) we have for the forces on m_1 due to m_2,

$$\frac{Y'}{y_2 - y_1} = \frac{Z'}{z_2 - z_1}, \tag{4}$$

and those on m_2 due to m_1 are equal and opposite. Hence the pair make a contribution to $\Sigma (yZ' - zY')$ equal to

$$(y_1 Z' - z_1 Y') - (y_2 Z' - z_2 Y') = 0, \tag{5}$$

by (4). Hence in (3) the reactions cancel in pairs and contribute zero to the total.

If we accept the atomic constitution of matter and suppose all atoms and electrons to act radially on one another this argument is valid. It has, however, seemed premature to many to accept it at the present stage. It is not obvious that such an ultimate analysis of the reactions is possible.

An alternative is to say that the internal forces depend on the body itself and not on outside agencies. Suppose then that the external forces are zero. If then the contributions to (2) and (3) from the internal forces were not zero, the body would begin to move of its own accord. For rigid bodies this does not happen, as a matter of experimental fact. But it seems wrong to generalize this to bodies under external forces and in a state of rotation. The internal forces are then certainly different from what they are in a stationary body under no external force, and this procedure gives us no ground for believing that the additional forces satisfy the rule.

It seems to me that the proper procedure is to recognize that the principle of d'Alembert has a moderate probability on account of the considerations on mutual influence of particles, and to investigate its consequences. If they are found to agree with experiment the principle becomes a

* If the particles are magnetic doublets this is not true.

scientific law in the usual sense. Now if we define the centroid of a body by the equations

$$(\Sigma m)\,\bar{x} = \Sigma mx; \quad (\Sigma m)\,\bar{y} = \Sigma my; \quad (\Sigma m)\,\bar{z} = \Sigma mz, \tag{6}$$

it follows from d'Alembert's principle that

$$(\Sigma m)\,\ddot{\bar{x}} = \Sigma X, \tag{7}$$

with two similar equations. The equations of motion that we have used hitherto are therefore satisfied by the coordinates of the centroid.

It remains to show that the centroid is actually fixed in the body. This is usually taken for granted, but it is not obvious that the centroid, which is so far merely a point whose coordinates are defined by (6), is also always at the same particle of the body. But if we consider the distance r_1 of the centroid from a given particle m_1 at (x_1, y_1, z_1) and denote other particles by m_l at (x_l, y_l, z_l) we have

$$(\Sigma m)\,(x_1 - \bar{x}) = (\Sigma m)\,x_1 - \Sigma mx$$

and

$$= \Sigma m_l\,(x_1 - x_l), \tag{8}$$

$$(\Sigma m)^2\,r_1^2 = \{\Sigma m_l\,(x_1 - x_l)\}^2 + \{\Sigma m_l\,(y_1 - y_l)\}^2 + \{\Sigma m_l\,(z_1 - z_l)\}^2. \tag{9}$$

The square terms on the right give

$$\Sigma m_l^2\,\{(x_1 - x_l)^2 + (y_1 - y_l)^2 + (z_1 - z_l)^2\} = \Sigma m_l^2 r_{1l}^2. \tag{10}$$

The product terms are of the form

$$2\Sigma m_l m_{l'}\,\{(x_1 - x_l)\,(x_1 - x_{l'}) + (y_1 - y_l)\,(y_1 - y_{l'}) + (z_1 - z_l)\,(z_1 - z_{l'})\} = 2\Sigma m_l m_{l'} r_{1l} r_{1l'} \cos(l1l'), \tag{11}$$

where by $(l1l')$ we mean the angle subtended at m_1 by the line joining the particles m_l and $m_{l'}$, and the summation is for all pairs of particles. But in a rigid body all distances and angles defined by pairs and triads of given particles are constant. Hence every term on the right of (10) and (11) is constant, and therefore r_1 is constant. Thus the centroid is at a constant distance from any particle of the body and therefore is fixed in the body itself.

Now we can denote the sum of the masses of the particles by M, which is the mass of the body as a whole. The equations (7) now reduce to the usual form

$$M\ddot{\bar{x}} = \Sigma X,$$

and so on. The other equations (3) can be shown, as is done in works on dynamics, to determine the motion of a rigid body about its centroid: that is to say, its rotation.

8·61. We have seen that the gravitational force on one body due to another is of the form fm_1m_2/r^2_{12}, acting along the line joining them. When the bodies are of finite size both the magnitude and direction of the force are somewhat vague, but we can make them precise also by considering the bodies as made up of particles. If then X is the force on a particle of mass m due to all other particles, we can write

$$(X, Y, Z) = m\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) V,$$

where

$$V = \Sigma \frac{fm'}{r},$$

where m' is the mass of another particle, r the distance between m and m', and the summation is for all the other particles. The function V has a definite value at all places inside or outside of bodies, and is called the *gravitation potential*. It can be applied to determine the quantities of the form ΣX and $\Sigma(yZ - zY)$ for any body, and hence to obtain differential equations for the motion of the body. For bodies that have not spherical symmetry the sums $\Sigma(yZ - zY)$ do not in general vanish, and consequently produce changes in the rotation. This result can be checked by appeal to the motion of the earth. The earth is not a sphere, but an oblate spheroid. The attractions of the sun and moon on it produce changes of the rotation of various types that can be predicted theoretically. The axis of figure describes a cone relative to the centroid of the earth, the axis of the cone being at right angles to the plane of the earth's orbit about the sun. This motion

is slow; the complete revolution takes 25,000 years. Superposed on it is an oscillation called a nutation, introduced by the fact that the moon's orbit about the earth is not always in the same plane. This causes the angular distance of the earth's pole of figure from the pole of the ecliptic to vary in a period of about 19 years, while the rate of its motion has a periodic variation in the same time. All the phenomena contain the factor $(C - A)/C$, where C and A are the earth's greatest and least moments of inertia. This factor is not determinable except from these phenomena. But the two components of the nutation depend on the mass of the moon, and not directly on that of the sun. The rate of the precession involves the masses of both bodies. Thus when we have observations of the precession and nutation we can use them to determine the ratio $(C - A)/C$ and the mass of the moon. It is found that the mass of the moon given by this method agrees with that found from the earth's monthly motion. Thus we have a quantitative check on the truth of d'Alembert's principle.

8·62. The important constant f has to be determined by direct observation of the attractive force between bodies of known mass at the earth's surface. It is found that the couple needed to twist a fine fibre of vitreous quartz through any angle is proportional to that angle. A bar with two lead spheres on the ends is suspended at its centroid from such a wire, and the period of the oscillation of the bar as it executes torsional oscillations is determined. The moment of inertia of the bar being known, this gives the couple exerted by the wire for any twist, in terms of c.g.s. units. Then the bar is allowed to take up its equilibrium position. Two large lead spheres are then arranged so that their attractions tend to twist the wire, and the head of the wire is then turned round until the torsion of the wire brings the bar back to its equilibrium position. The amount of turn required determines the magnitude of the attractions of the spheres and hence the constant f.

The equations of motion can be put into a form depending as directly on fundamental concepts as the conservation of energy. Suppose that the co-ordinates of a particle of a system of mass m moving according to these equations are (x, y, z). Then (x, y, z) are definite functions of the time. Take any three other functions of the time $(\delta x, \delta y, \delta z)$, restricted only to be differentiable. Then the equations of motion are equivalent to

$$m(\ddot{x}\delta x + \ddot{y}\delta y + \ddot{z}\delta z) = X\delta x + Y\delta y + Z\delta z. \quad (1)$$

Suppose these equations added for all particles of the system, and the result integrated from time t_0 to time t_1. Then

$$\int_{t_0}^{t_1} \Sigma m(\ddot{x}\delta x + \ddot{y}\delta y + \ddot{z}\delta z)\,dt = \int_{t_0}^{t_1} \Sigma(X\delta x + Y\delta y + Z\delta z)\,dt. \quad (2)$$

Now imagine the system to be moved from time t_0 to t_1 in such a way that at time t the co-ordinates of the particle m are $(x + \delta x, y + \delta y, z + \delta z)$. We may call $x + \delta x$ a *varied* co-ordinate and $\dot{x} + \delta\dot{x}$ a varied component of velocity. Then

$$\delta\dot{x} = (\dot{x} + \delta\dot{x}) - \dot{x} = \frac{d}{dt}(x + \delta x) - \frac{dx}{dt} = \frac{d}{dt}\delta x. \quad (3)$$

The left side of (2) is

$$= \left[\Sigma m(\dot{x}\delta x + \dot{y}\delta y + \dot{z}\delta z)\right]_{t_0}^{t_1} - \int_{t_0}^{t_1} \Sigma m\left(\dot{x}\frac{d}{dt}\delta x + \dot{y}\frac{d}{dt}\delta y + \dot{z}\frac{d}{dt}\delta z\right)dt. \quad (4)$$

But $$\dot{x}\frac{d}{dt}\delta x = \dot{x}\delta\dot{x} = \tfrac{1}{2}(\dot{x} + \delta\dot{x})^2 - \dot{x}^2 - \delta\dot{x}^2, \quad (5)$$

so that if $$T = \Sigma\tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2), \quad (6)$$

$$\Sigma m\left(\dot{x}\frac{d}{dt}\delta x + \dot{y}\frac{d}{dt}\delta y + \dot{z}\frac{d}{dt}\delta z\right) = \delta T - \Sigma m(\delta\dot{x}^2 + \delta\dot{y}^2 + \delta\dot{z}^2). \quad (7)$$

If $(\delta x, \delta y, \delta z)$ vanish at times t_0 and t_1, so that the varied co-ordinates begin and end at the same values as the actual ones, the first term in (4) is zero. Then

$$\int_{t_0}^{t_1} \{\delta T + \Sigma (X\delta x + Y\delta y + Z\delta z)\}\, dt = O(\delta\dot{x}, \delta\dot{y}, \delta\dot{z})^2. \quad (8)$$

Now the forces in the system may be conservative, so that when the system is displaced from one position to another, by any route, the forces do the same amount of work. Then there is a function U depending only on the relative positions of parts of the system, such that

$$\Sigma (X\delta x + Y\delta y + Z\delta z) = \delta U + O(\delta x, \delta y, \delta z)^2. \quad (9)$$

Then finally

$$\int_{t_0}^{t_1} \delta (T + U)\, dt = O(\delta x, \delta y, \delta z, \delta\dot{x}, \delta\dot{y}, \delta\dot{z})^2. \quad (10)$$

If we define a function

$$S = \int_{t_0}^{t_1} (T + U)\, dt, \quad (11)$$

then δS for small variations in the path is of the *second* order in those variations. This is Hamilton's principle.

If some of the particles constitute a rigid body and $(\delta x, \delta y, \delta z)$ are such as not to alter the distances between them, it can be shown that d'Alembert's principle implies that the internal forces contribute nothing to $\Sigma (X\delta x + Y\delta y + Z\delta z)$. We can therefore restrict $(\delta x, \delta y, \delta z)$ to depend only on the translations and rotations of rigid bodies and omit the contributions from the internal forces.

CHAPTER IX

LIGHT AND RELATIVITY

It did not last: the Devil, howling Ho!
Let Einstein be! restored the *status quo*.

J. C. SQUIRE

9·1. We have seen that the truth of the equations of dynamical astronomy in the form

$$\ddot{x}_l = -\Sigma \mu_m \frac{x_l - x_m}{r^3_{lm}},$$

requires that the axes shall have no rotation and that the origin shall have a uniform velocity with respect to the centroid of the universe. We have also supposed that at each instant of time each particle of the system has definite coordinates with respect to these axes. With regard to the time some further discussion is necessary. At first astronomers thought that the time in these equations was the time of observation. But this was found to be incorrect by Römer. When the periods of revolution of Jupiter's satellites were found by observation of their eclipses, transits, and occultations when Jupiter was near opposition, the results were used to predict these phenomena when Jupiter was situated otherwise; and errors ranging up to a quarter of an hour were found. Römer gave the correct explanation, namely that the time in the equations of dynamics is not the time of observation, but the time when the event under discussion actually happened, and that in a visual observation the time of observation is later because it takes light a finite time to travel from the object observed to the observer. Jupiter in opposition is nearer to us than at other times, and therefore light takes a shorter time to travel to us. Events at that time therefore do not suffer such a great delay in being observed as when Jupiter is at greater distances. The delay corresponds

to a time of passage across the earth's orbit of about 16 minutes. When the times were corrected to allow for this effect the anomalies disappeared,

Direct measurement of the velocity of light near the earth's surface was carried out by Fizeau and Foucault. The methods depend on the principle of sending out an intermittent beam of light to a distant object, where it is reflected, and observing the time that elapses between the flashes going out and coming back. It was found, as expected, that this time was proportional to the distance traversed. With the best modern methods, Michelson has obtained a velocity of 299,796 ± 4 km./sec. The experimental determinations agree with that found from the observations of Jupiter's satellites within the uncertainty of the latter.

When the source of light has a velocity of its own, various possibilities arise with regard to the effect of this velocity on the velocity of light. If light consisted of a stream of corpuscles, it might be expected that the velocity of the source would be added vectorially on to the velocity of the emitted light. But if the velocity of light is a fundamental physical constant we might expect that when light gets away from the source it settles down to move with its standard velocity and forgets about its source. The velocity of light is so great that a small alteration of it, such as the motion of the source can introduce, would not affect observations within the solar system, but in double stars the effect might be sensible. In the eclipsing binary Algol, for instance, the orbital velocity appears to be about 240 km./sec., or $0{\cdot}8 \times 10^{-3}$ of the velocity of light. The distance of the star is about 35 parsecs or 10^{15} km., so that light from it ordinarily takes 3×10^{9} seconds to reach us. The effect on the time due to a change of 0·8 parts in a thousand in the velocity of light would therefore be $2{\cdot}4 \times 10^{6}$ seconds or 26 days. The whole period of revolution of the star is under 3 days. Thus light from the fainter component when it is approaching us would reach us *before* the light that leaves us the next time it begins to recede from us,

and the apparent variation with time of the position of the secondary would be completely upset*. The secondary, as it happens, is not visible separately, being too close to the primary, and is known to us principally from its regular eclipsing of a portion of the primary's surface. Application of the test to Algol may therefore be impossible; it is mentioned here merely as an illustration.

9·2. This theory of the velocity of light from a moving source has never, as a matter of fact, been taken seriously. We know from the phenomenon of interference that light is a wave motion. The velocity of waves is a matter of the physical properties of the region they are traversing; once away from the source they look after themselves.

Now consider a moving system consisting of a source of light, with a mirror at distance l away from it in the direction of the velocity. The source and the mirror are both moving with velocity v. Then, on the natural way of looking at the matter, light leaving the source has a velocity c and is gaining on the mirror with relative velocity $c - v$. Hence it will overtake the mirror in time $l/(c - v)$. After reflexion it is moving with velocity c again, but towards the source, and has a relative velocity $c + v$. Hence it returns to the source in time $l/(c + v)$, and the total time taken is

$$\frac{l}{c - v} + \frac{l}{c + v} = \frac{2cl}{c^2 - v^2}.$$

Now suppose that the direction of the mirror from the source is at right angles to the velocity of the source. Then after the light leaves the source, the mirror and source both go on moving with velocity v. If t is the total time of transit from the source to the mirror and back the source has meanwhile travelled a distance vt, and if the light on return reaches the new position of the source it has a component of velocity v in the direction of motion of the latter. Hence its transverse

* The data are taken, roughly, from Eddington, *The Internal Constitution of the Stars*, p. 209.

component of velocity is $(c^2 - v^2)^{\frac{1}{2}}$. But the total distance travelled transversely is $2l$. Hence the time taken is $\frac{2l}{(c^2-v^2)^{\frac{1}{2}}}$. This is shorter than the time in the former case, in the ratio $(1-v^2/c^2)^{\frac{1}{2}}$. If then we can arrange for the two specimens of light to leave the source at the same time and for the distances of the two mirrors to be equal, the light that has travelled transversely will arrive back first, and therefore in a different phase. If the difference of phase is great enough interference will take place.

This experiment was carried out by Michelson and Morley, and has since been repeated by various other investigators. The two specimens of light were produced by a mirror silvered to semitransparency and inclined at 45° to the original beam from a lamp. Half the light went through to the mirror in the direction of the velocity of the system; the other half went transversely to the other mirror. On reflexion to the semitransparent mirror, the latter transmitted half of one beam and reflected half the other, the resulting beams now travelling in the same direction. The distances were made nearly equal. The whole apparatus was then turned through a right angle, so that the time of transit of one beam should be increased and that of the other diminished, and if interference did not occur in the first case it should occur in the second.

It was actually found that the rotation of the apparatus through a right angle made no difference. If two waves took the same time to travel backwards and forwards with one setting of the apparatus, they did so again with any other setting. The velocity of the system in this experiment was the resultant of the velocity of the earth in its orbit and that of the sun relative, one supposes, to the centroid of the universe. The latter may be supposed constant; the former is reversed every six months. It might perhaps happen that the two cancelled in one position of the earth with respect to the sun; but actually the result was the same at whatever time of the year the experiment was carried out.

9·21. The null result of the experiment showed that there was something wrong with the premisses. The ratio of the time of transmission to the distance was the same whatever the orientation of the distance with respect to the velocity of the system as a whole. The distances, for this purpose, are the distances between the points of reflexion of the light, as measured when there is no relative motion, and are therefore to be understood as in mensuration. Distance being taken in this sense, it appeared that the apparent velocity of light, measured as the distance travelled divided by the total time of passage there and back, was the same in any direction. The expression of this result is that if (x, y, z) are the measurable co-ordinates, with respect to an origin, of the place where a light-wave is at time t, then

$$\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 = c^2, \qquad (1)$$

irrespective of the motion of the origin and the direction of the axes. If then we take another origin and use analogous variables (x', y', z', t') we shall have also

$$\left(\frac{dx'}{dt'}\right)^2 + \left(\frac{dy'}{dt'}\right)^2 + \left(\frac{dz'}{dt'}\right)^2 = c^2. \qquad (2)$$

Either of these forms implies the other. If we write

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2;$$
$$ds'^2 = c^2dt'^2 - dx'^2 - dy'^2 - dz'^2, \qquad (3)$$

then if either of ds and ds' is zero, the other is also zero. It follows that for all positions and times

$$ds' = k\,ds, \qquad (4)$$

where k may be a function of (x, y, z, t). When (x, y, z, t) are determined, (x', y', z', t') are also determinate, so that the variables referred to one system of reference are definite functions of those referred to the other. Thus k cannot involve the velocities.

Now suppose that (x, y, z, t) and (x', y', z', t') are both such

systems of reference that the equations of dynamics hold with regard to them. Then if $\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}$ for a particle are constant, the particle is under no forces, and therefore $\frac{dx'}{dt'}, \frac{dy'}{dt'}, \frac{dz'}{dt'}$, are also constant. Also $\frac{ds}{dt}$ and $\frac{ds'}{dt'}$ are constant. Therefore the two sets of equations

$$\frac{dx}{ds}, \frac{dy}{ds}, \frac{dz}{ds}, \frac{dt}{ds} = \text{constants}, \tag{5}$$

and

$$\frac{dx'}{ds'}, \frac{dy'}{ds'}, \frac{dz'}{ds'}, \frac{dt'}{ds'} = \text{constants}, \tag{6}$$

are equivalent. Now consider a particle to move from one given position at time t_0 to another at time t_1. Then if (x, y, z) are given as functions of t_0 during the transit, $\int_{t_0}^{t_1} \frac{ds}{dt}\, dt$ has a definite value. If we choose a slightly different path, the change of this integral is $\delta \int ds$. We can show that the conditions (5) are just the conditions that $\delta \int ds$ shall be of the *second* order in the variations of (x, y, z). Thus the equations of motion of a particle under no forces are equivalent to the statement that $\delta \int ds$ is of the second order. Similarly they are equivalent to the statement that $\delta \int ds'$ is of the second order. Hence if a path is such that $\delta \int ds$ is of the second order, so is $\delta \int k\, ds$.

It follows that k is constant. For if k depended on (x, y, z) we could take a path near the original one, but always on the side of it where k is greater than on the actual path. Then $\delta \int ds$ would be of the second order, but $\delta \int k\, ds$ would be of the first order. If k depended on t, we could take the same path but alter the rate of travel so that t has slightly different values for the same (x, y, z). We could arrange for

$$(dx^2 + dy^2 + dz^2)/dt^2$$

to be increased when k is large and decreased when k is small. Then we get a first order change in $\int k\, ds$. Thus k must

be independent of (x, y, z, t). It can be seen that it is unity. For if the origin of (x', y', z', t') has a velocity with regard to that of (x, y, z, t), we can turn both sets of axes of (x, y, z), (x', y', z') round so that those of x and x' are in the direction of this velocity. A bar of length l at right angles to this direction in one system has length l in the other system, and as we go along it $dx = dx' = 0$, $dt = dt' = 0$, so that

$$l = kl, \tag{7}$$

and therefore

$$k = 1. \tag{8}$$

Thus

$$dx^2 + dy^2 + dz^2 - c^2dt^2 = dx'^2 + dy'^2 + dz'^2 - c^2dt'^2. \tag{9}$$

Also we can write

$$\frac{dx'}{ds'} = \frac{dx'}{ds} = \frac{dx}{ds}\frac{\partial x'}{\partial x} + \frac{dy}{ds}\frac{\partial x'}{\partial y} + \frac{dz}{ds}\frac{\partial x'}{\partial z} + \frac{dt}{ds}\frac{\partial x'}{\partial t}, \tag{10}$$

with similar equations. Now for a particle in uniform motion dx'/ds', dx/ds and similar expressions are constants, but these constants are different for different particles. Hence this can be true in general only if $\frac{\partial x'}{\partial x}, \frac{\partial x'}{\partial y}, \frac{\partial x'}{\partial z}, \frac{\partial x'}{\partial t}$ are also constants. Therefore (x', y', z', t') are linear functions of (x, y, z, t).

Now again take the axes in the direction of relative motion of the origin. Then we can arrange the directions of the y' and z' axes so that:

$$\left.\begin{array}{lll} \text{If} & x = Vt, & x' = 0. \\ \text{If} & y = 0, & y' = 0. \\ \text{If} & z = 0, & z' = 0. \end{array}\right\} \tag{11}$$

Hence the relations between the co-ordinates are of the form

$$\left.\begin{array}{l} x' = \beta(x - Vt) \\ y' = \gamma y \\ z' = \delta z \\ t' = \alpha_1 x + \beta_1 y + \gamma_1 z + \delta_1 t + \epsilon. \end{array}\right\} \tag{12}$$

Now substitute in (9). We have

$$\begin{aligned}&\beta^2(dx - V\,dt)^2 + \gamma^2 dy^2 + \delta^2 dz^2\\ &\quad - c^2(\alpha_1 dx + \beta_1 dy + \gamma_1 dz + \delta_1 dt)^2\\ &\quad \equiv dx^2 + dy^2 + dz^2 - c^2 dt^2. \qquad (13)\end{aligned}$$

On equating coefficients of $dx\,dy$, $dx\,dz$, $dy\,dt$, $dz\,dt$ we have

$$\alpha_1\beta_1 = \alpha_1\gamma_1 = \beta_1\delta_1 = \gamma_1\delta_1 = 0. \qquad (14)$$

Hence either $$\alpha_1 = \delta_1 = 0, \qquad (15)$$

or $$\beta_1 = \gamma_1 = 0. \qquad (16)$$

If (15) is true, the coefficients of dx^2 and dt^2 give

$$\beta^2 = 1\,; \quad -\beta^2 V^2 = -c^2, \qquad (17)$$

so that this condition cannot arise unless the relative velocity of the origins is the velocity of light. The alternative (16) gives, from the coefficients of dy^2 and dz^2,

$$\gamma = \pm 1, \quad \delta = \pm 1\,; \qquad (18)$$

and we may agree to take the axes of y' and z' so that the positive signs are applicable. From the coefficients of dx^2, $dx\,dt$, and dt^2,

$$\beta^2 - c^2\alpha_1^2 = 1\,; \quad \beta^2 V + \alpha_1\delta_1 c^2 = 0\,;$$
$$\beta^2 V^2 - c^2\delta_1^2 = -c^2. \qquad (19)$$

Eliminating β and V we have

$$(1 + c^2\alpha_1^2)(\delta_1^2 - 1) = \alpha_1^2\delta_1^2 c^2, \qquad (20)$$

whence $$\delta_1^2 = 1 + \alpha_1^2 c^2 = \beta^2, \qquad (21)$$

and from the last of (19)

$$\beta^2 = \frac{c^2}{c^2 - V^2}. \qquad (22)$$

If x' and x are to increase in the same direction we must take the positive sign. Similarly we shall take δ_1 positive. Then the second of (19) gives

$$\alpha_1 = -\beta V/c, \qquad (23)$$

and finally

$$x' = \beta(x - Vt)\,; \quad y' = y\,; \quad z' = z\,; \quad t' = \epsilon + \beta\left(t - \frac{Vx}{c^2}\right), \qquad (24)$$

where $$\beta = (1 - V^2/c^2)^{-\frac{1}{2}}, \qquad (25)$$

and ϵ is an additive constant depending on the instant we measure t' from. If we choose this suitably, ϵ can be dropped.

If we solve (24) for x, y, z, t, we obtain

$$x=\beta(x'+Vt'); \quad y=y'; \quad z=z'; \quad t=\beta\left(t'-\epsilon+\frac{Vx'}{c^2}\right). \quad (26)$$

The equations (24) were first given in full by Sir Joseph Larmor, who showed that a transformation of this type leaves the form of the equations of the electromagnetic field unaltered. In the hands of Einstein they became the basis of the special theory of relativity.

It appears from the first of (24) that $(V, 0, 0)$ is the velocity of the origin of the second system with reference to the first, and from the first of (26) that $(-V, 0, 0)$ is the velocity of the origin of the first system with reference to the second.

9·22. Now consider two events specified in the first system by (x_1, y_1, z_1, t_1), (x_2, y_2, z_2, t_2) and in the second by corresponding letters with accents. Then

$$x_2'-x_1'=\beta\{x_2-x_1-V(t_2-t_1)\}, \quad (1)$$

$$y_2'-y_1'=y_2-y_1, \quad (2)$$

$$z_2'-z_1'=z_2-z_1, \quad (3)$$

$$t_2'-t_1'=\beta\{t_2-t_1-V(x_2-x_1)/c^2\}. \quad (4)$$

Hence distances perpendicular to the direction of relative motion are the same in the two systems.

Suppose the events are simultaneous in the second system, so that $t_2'=t_1'$. Then from (4)

$$t_2-t_1=V(x_2-x_1)/c^2, \quad (5)$$

and on substitution in (1)

$$\begin{aligned} x_2'-x_1' &= \beta(1-V^2/c^2)(x_2-x_1) \\ &= (1-V^2/c^2)^{\frac{1}{2}}(x_2-x_1). \end{aligned} \quad (6)$$

If then x_2-x_1 is independent of the time, $x_2'-x_1'$ is constant and less than x_2-x_1 in a definite ratio. If on the other hand $t_2=t_1$, we shall find

$$x_2-x_1=(1-V^2/c^2)^{\frac{1}{2}}(x_2'-x_1'). \quad (7)$$

Now when we measure a distance on a moving object we compare the positions of the ends of the object simultaneously with those of points with no motion relative to our axes. The equality of t_2' and t_1' is essential to the attribution of any meaning to $x_2' - x_1'$ in terms of distance when the co-ordinates themselves are varying with time. If we have an object whose length in the x direction in the first system is independent of the time in the first system, then its length in the second system is less than its length in the first in the ratio $(1 - v^2/c^2)^{\frac{1}{2}}$, and conversely. This apparent contraction of a moving object in the direction of motion is known as the Fitzgerald contraction, and depends, as we see from (5), on the fact that two observers in relative motion differ in their ideas of what events are simultaneous on account of the finiteness of the velocity of light.

9·23. We can now identify the time of a distant event in terms of light signals. For if a mirror is at a fixed distance l, then light takes a time l/c to reach it, and a further time l/c to return. Thus the time when reflexion occurs is the mean of those when the wave leaves the source and returns to it. This result is irrespective of the velocity of the system, and is not true on the older theory, where the times of the outgoing and returning waves were liable to differ, as we saw in discussing the Michelson-Morley experiment.

The chief difficulty usually felt in relation to the modern theory of relativity is precisely in connexion with the result that events that are simultaneous to one observer are not simultaneous to another. But this difficulty arises in reality at a much earlier stage than the transformation (24). If the time means the time of observation, then the observations of Jupiter's satellites prove that the equations of dynamics are untrue, and conversely, if we are to retain the equations of dynamics, the time of an event is not the time of observation. We must therefore have a rule to enable us to infer the one from the other, of such a character as will keep the equations

of dynamics true. The only rule that will satisfy the criteria that we have already is (24). The time of observation being different from the time of the event, and depending on the position of the observer, as a matter of observation, the whole conception of simultaneity required rediscussion from the start. In any case the times of different observers' observations of the same event differ by quantities of the order of the differences of r/c, where r is the distance travelled by the light. In the new theory we have obtained a time of the event itself, which varies for different observers by quantities of the order of Vr/c^2. But V/c is in general small; thus the deviations between different observers from agreement about time-intervals are of the second order of small quantities instead of the first. The objection is in fact a straining at the gnat, while swallowing the camel without even noticing the existence of the larger animal.

9·24. Now consider a point moving with velocities (u, v, w) with reference to the system (x, y, z, t). Then its velocities with reference to the system (x', y', z', t') are

$$u' = \frac{dx'}{dt'} = \frac{\beta(dx - Vdt)}{\beta(dt - Vdx/c^2)} = \frac{u - V}{1 - uV/c^2}, \qquad (1)$$

$$v' = \frac{dy'}{dt'} = \frac{dy}{\beta(dt - Vdx/c^2)} = \frac{v}{\beta(1 - uV/c^2)}, \qquad (2)$$

$$w' = \frac{w}{\beta(1 - uV/c^2)}. \qquad (3)$$

We notice that if $(u, v, w) = (c, 0, 0)$, then $(u', v', w') = (c, 0, 0)$ whatever V may be. If $(u, v, w) = (0, c, 0)$, then $(u', v', w') = (-V, c/\beta, 0)$ and

$$u'^2 + v'^2 + w'^2 = V^2 + c^2(1 - V^2/c^2) = c^2.$$

These results are of course particular cases of our fundamental rule that the velocity of light is the same however the observer is moving.

We notice a curious phenomenon if V should happen to be greater than c, the velocity of light. Imagine a particle moving with velocities (u, v, w), and consider its velocities with respect to an origin moving with velocity V. The quantity β is imaginary if $V/c > 1$. Hence v' and w' are imaginary, and the particle could have real co-ordinates at only one instant; for the rest of time its co-ordinates are imaginary—which is as much as to say that the particle is imaginary. There seems to be no inherent contradiction in the idea of velocities greater than that of light: but if we consider as our universe all particles moving with velocities less than c with respect to ourselves, then any particle with a velocity greater than c with respect to ourselves has a velocity greater than c with respect to any other particle of our universe. The world could then be classified into universes, such that no particle in any one universe could be perceptible for more than a fleeting instant from a different universe.

9·25. We now consider other observable consequences of the transformation. Consider a source of light sending out waves of period $2\pi/\gamma$ along the axis of x. Then the disturbance at any distance contains a factor such as

$$\phi = A \sin \gamma (t - x/c). \qquad (1)$$

Now consider an observer with a velocity V along the x axis. Using (26) we have

$$\phi = A \sin \gamma\beta \left(t' + \frac{Vx'}{c^2} - \frac{x'}{c} - \frac{Vt'}{c}\right) = A \sin \gamma\beta \left(1 - \frac{V}{c}\right)\left(t' - \frac{x'}{c}\right)$$

$$= A \sin \gamma' \left(t' - \frac{x'}{c}\right), \qquad (2)$$

where

$$\gamma' = \gamma\beta (1 - V/c), \qquad (3)$$

so that the period of the disturbance reaching the observer is longer than that estimated by a stationary observer in the ratio $\beta (1 - V/c) : 1$. In practice V/c is always small and β

indistinguishable from unity. But the factor $1 - V/c$ produces an apparent lengthening of the wave-length of a given spectral line for a receding star, and a shortening of it for an approaching one. This is the Doppler effect, and is measurable. It leads to estimates of the radial velocities in double stars which agree with those inferred from the transverse movements, and has many other astronomical applications.

9·26. Now suppose that an observer in the (x, y, z, t) system sees a star in the direction (l, m, n). Then the velocity components of the light from the star are

$$(u, v, w) = -(lc, mc, nc). \tag{1}$$

To an observer in the (x', y', z', t') system the apparent direction is (l', m', n') and velocity components are, by 9·24,

$$-l'c = u' = \frac{-lc - V}{1 + lV/c}; \quad -m'c = v' = -\frac{mc}{\beta(1 + lV/c)};$$

$$-n'c = w' = -\frac{nc}{\beta(1 + lV/c)}. \tag{2}$$

Thus $m'/n' = m/n$, and the directions (l, m, n) and (l', m', n') lie in a plane including the axis of x. If

$$l = \cos\theta; \quad l' = \cos\theta', \tag{3}$$

$$l' - l = \frac{1 + V/lc}{1 + lV/c} - 1 = \frac{1 - l^2}{l}\frac{V}{c}\frac{1}{1 + lV/c}, \tag{4}$$

or, if we neglect the square of V/c,

$$\theta' - \theta = -\frac{V}{c}\tan\theta. \tag{5}$$

Thus the apparent direction of the star is displaced towards the direction of the relative motion of the second observer by an amount given by (5). This is the phenomenon of *aberration*. On account of the earth's orbital motion its velocity relative to the sun varies in the course of a year, and therefore pro-

duces periodic variations in the apparent directions of the stars. These are the same for all stars in the same part of the sky, and are well known to astronomers.

9·27. Now consider light emitted from a source and entering water moving with velocity V in the direction of the beam. The velocity of light relative to the water is c/μ, where μ is the refractive index. The velocity of the source relative to the water is $-V$. Using 9·24 (1) to get the velocity of the water relative to the source, we get

$$u' = \frac{c/\mu + V}{1 + (c/\mu)V/c^2} = \frac{c}{\mu} + V\left(1 - \frac{1}{\mu^2}\right) + O\left(\frac{V^2}{c^2}\right). \quad (1)$$

This is tested in Fizeau's experiment. Water travels in a closed pipe so that when the light is travelling outwards to a distant mirror the water is moving with it, and the reflected beam travels with the return current, so that the effect of V is to increase u' on both the outward and the return journeys. Another beam is sent round the other way, so that its apparent velocity is reduced by the motion of the water. The two beams are recombined on return and the difference in the times of travel measured by a method of interference. It was found in a repetition of the experiment by Michelson and Morley that the observed value* of the coefficient of V was $0{\cdot}442 \pm 0{\cdot}02$. The value calculated from the refractive index μ was $0{\cdot}438$; but this became $0{\cdot}451$ when a refinement allowing for dispersion was made. The agreement is within the error of observation. A further repetition by Zeeman gave almost perfect agreement.

The result of Fizeau's experiment is extremely important. If the velocity of light in a moving medium was the sum of the ordinary velocity and the velocity of the medium, the coefficient of V would have been 1. If the velocity was independent of that of the medium it would have been 0.

* A numerical correction due to Cunningham, *Relativity and the Electron Theory*, 1920, has been used.

The experiment excludes both of these alternatives. It shows also that the actual coefficient agrees with that calculated from (1); and the term in $1/\mu^2$, if we trace it back, is found to come from the uV/c^2 in the denominator of 9·24 (1), which came in turn from the term in Vx/c^2 in the expression for t'. Thus it gives a direct check on this term, which is the very one in the fundamental transformation that has been most subject to dispute.

9·3. The foregoing theory is usually known as Einstein's special theory of relativity, though the use of the word *relativity* as if it expressed a novel feature is really incorrect. The relativity of the equations of dynamics, in the sense that they are true whatever unaccelerated non-rotating axes we use, had been known since Newton. The need for Einstein's theory arose from two facts about light: first, that it has a finite velocity, and second, that this velocity is independent of the motion of the observer. If light had travelled with an infinite velocity we could have identified the time of the event with the time of observation, and there would have been no further trouble. But this ceased to be a serious possibility when Römer made his discovery about Jupiter's satellites; the time in the equations of dynamics is not the time when the observations are made. The modern problem is not the discovery of relativity, but to retain relativity without introducing inconsistency with what we know about light. Even with the modification we have made so far, in the relations between the co-ordinates and time in different systems of reference, the equations of dynamics lose their relativistic form.

9·31. Consider two bodies moving according to the equations

$$m_1\ddot{x}_1 = -m_2\ddot{x}_2 = fm_1m_2\frac{x_2 - x_1}{r^3}, \qquad (1)$$

and so on. The co-ordinates and time are those of an unaccelerated observer. Now imagine another observer with

velocities (V, 0, 0) with reference to the first. Applying the transformation of 9·24 we get

$$\frac{d^2x'}{dt'^2} = \frac{1}{\beta^3(1-uV/c^2)^3}\frac{d^2x}{dt^2}, \tag{2}$$

$$\frac{d^2y'}{dt'^2} = \frac{1}{\beta(1-uV/c^2)^3}\frac{d^2y}{dt^2} + \frac{V}{c^2\beta(1-uV/c^2)^3}\left(\frac{dy}{dt}\frac{d^2x}{dt^2} - \frac{dx}{dt}\frac{d^2y}{dt^2}\right), \tag{3}$$

with an analogous expression for d^2z'/dt'^2.

It is clear from these equations that in the new system $\frac{d^2x_1'}{dt'^2}$ and $\frac{d^2x_2'}{dt'^2}$ cannot be in a constant ratio. For in (2), u appears explicitly, and is different for the two bodies and variable for both. The special theory is not relativistic when applied to the equations of dynamics, except for unaccelerated particles.

We may notice, however, that with ordinary velocities ds^2/c^2dt^2 is nearly 1, and $ds' = ds$. We might try then to modify the equations by replacing d/dt by cd/ds. Then

$$\frac{dx'}{ds'} = \frac{d}{ds}\beta(x-Vt) = \beta\left(\frac{dx}{ds} - V\frac{dt}{ds}\right); \quad \frac{dy'}{ds'} = \frac{dy}{ds}; \quad \frac{dz'}{ds'} = \frac{dz}{ds}; \tag{4}$$

$$\frac{d^2x'}{ds'^2} = \beta\left(\frac{d^2x}{ds^2} - V\frac{d^2t}{ds^2}\right); \quad \frac{d^2y'}{ds'^2} = \frac{d^2y}{ds^2}; \quad \frac{d^2z'}{ds'^2} = \frac{d^2z}{ds^2}. \tag{5}$$

But
$$c^2\left(\frac{dt}{ds}\right)^2 = 1 - \left(\frac{dx}{ds}\right)^2 - \left(\frac{dy}{ds}\right)^2 - \left(\frac{dz}{ds}\right)^2, \tag{6}$$

$$c^2\frac{dt}{ds}\frac{d^2t}{ds^2} = -\frac{dx}{ds}\frac{d^2x}{ds^2} - \frac{dy}{ds}\frac{d^2y}{ds^2} - \frac{dz}{ds}\frac{d^2z}{ds^2}. \tag{7}$$

Thus d^2x'/ds'^2 depends not only on the accelerations in the original frame of reference but on the velocities, and the velocities are variable and different for the two bodies. Hence the values of d^2x'/ds'^2 for the two bodies are still not in a constant ratio, and the equations of dynamics do not satisfy the principle of relativity.

9·4. We saw that the special theory depended on two postulates: a particle moving with uniform velocity with respect to one system of reference has uniform velocity with respect to any other system; and the velocity of light is the same in any system. We saw also that these propositions can be expressed by saying that $ds = ds'$, that the path of an unaccelerated particle is specified by the equation $\delta \int ds = 0$ to the first order, and that the path of a light wave is the limit of that of a particle when the velocity approaches c. These have a very general form. Now we saw that we could put the equations of dynamics in a very general form

$$\delta \int_{t_0}^{t_1} \{\Sigma \tfrac{1}{2} m (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) + U\}\, dt = 0, \qquad (1)$$

to the first order. For particles under no forces $U = 0$. But

$$\int ds = \int \frac{ds}{dt}\, dt = c \int \left(1 - \frac{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}{c^2}\right)^{\frac{1}{2}} dt$$
$$= c \int \{1 - \tfrac{1}{2} (\dot{x}^2 + \dot{y}^2 + \dot{z}^2)/c^2 + O(c^{-4})\}\, dt. \qquad (2)$$

Now if we do not vary the values of (x, y, z, t) at the limits the first term of (2) is just $c(t_1 - t_0)$ and its variation is zero. The second term, apart from a constant factor, leads to an equation of the same form as (1) takes for an unaccelerated particle. This strongly suggests that there is an analogy between Hamilton's principle and the stationary property of $\int ds$. If in fact we consider

$$\int \left\{\Sigma m (c^2 dt^2 - dx^2 - dy^2 - dz^2)^{\frac{1}{2}} - \frac{U}{c}\, dt\right\}, \qquad (3)$$

we have an integral that behaves in the proper way when U is zero, and yields Hamilton's principle as an approximation, with errors of order c^{-2}, when U is variable. Alternatively, if we introduce V_l, the gravitation potential at the particle m_l, we have

$$U = \Sigma \frac{f m_l m_m}{r} = \tfrac{1}{2} \Sigma_l m_l V_l, \qquad (4)$$

since $\Sigma_l m_l V_l$ takes each pair of particles twice. Then

$$\Sigma m_l (c^2 dt_l^2 - dx_l^2 - dy_l^2 - dz_l^2 - 2V_l dt_l^2)^{\frac{1}{2}}$$
$$= c \Sigma m_l \left\{ 1 - \frac{V_l}{c^2} - \frac{\dot{x}_l^2 + \dot{y}_l^2 + \dot{z}_l^2}{2c^2} + O(c^{-4}) \right\} dt, \quad (5)$$

so that if we redefine ds_l by the equation

$$ds_l^2 = (c^2 - 2V_l) dt_l^2 - dx_l^2 - dy_l^2 - dz_l^2, \quad (6)$$

we can sum up our present knowledge in the form

$$\delta \Sigma \int m_l ds_l = 0, \quad (7)$$

to the first order.

Clearly an infinite number of such hypotheses would satisfy our present data equally well; for all that is necessary is that the integrand should reduce, at a great distance from attracting bodies, to the *ds* of the special theory, and that near attracting bodies the second approximation should be of the form

$$c \Sigma m_l \left\{ 1 - \frac{1}{2c^2} (\dot{x}^2 + \dot{y}^2 + \dot{z}^2 + 2V) \right\} dt. \quad (8)$$

Any terms of order $(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)^2/c^4$ or V^2/c^4 could be included in the coefficient of *dt* without disturbing our present knowledge.

At this stage there are two possible lines of progress. One is that actually adopted by Einstein, which led to the general theory of relativity. We notice that the Newtonian equations have a form that is unaffected by a uniform velocity of the origin, while the properties of light and freely moving particles at a distance from matter can be put in a simple form depending only on the *ds* of the special theory, which is independent of the choice of origin. This property, that the fundamental equations are independent of the velocity of the origin (and of course the directions of the axes, so long as they are not rotating), is of a very simple and general character, and therefore has a moderate prior probability. Its verification to a considerable order of accuracy by the phenomena considered in the special theory and by the laws of Newtonian dynamics therefore establishes a high probability that it is

true exactly and in general. We may therefore take it as a fundamental postulate and develop its consequences. If these turn out to be verified its probability will approach certainty.

The other line of attack is to begin with the observed phenomena and find what is the simplest law that fits them. It is found that this law has relativistic properties, and affords justification for trying to push the principle of relativity still further.

9·41. Starting from our recognition in the special theory of the fundamental importance of ds, we see from (7) that there is a possibility of retaining it in a gravitational field, provided we modify the coefficients slightly. But if ds is to have such an importance it must be the same for all observers, and it is easy to see that this casts our whole scheme of Cartesian co-ordinates and time into the melting-pot. Imagining the coefficients to have been modified suitably, we must suppose, as our obvious generalization from the special theory applicable in the absence of a gravitational field, that the motion of particles in a gravitational field is such that $\int ds$ is stationary for small variations in the path, and that the path of a light wave is the limit of the path of a particle when it is such that $ds = 0$ between any two consecutive points on the path. But in that case, since gravity appears explicitly in ds, light is affected by gravity, light rays may be curved in a gravitational field, and our test of collinearity among distant objects breaks down; our laws relating distances then become approximations and do not hold exactly. We might try to save them by saying that in a gravitational field the ds suitable for light still has constant coefficients, so that light still travels in straight lines, but that the form suitable for material particles does involve the gravitational field. But at the present time this possibility is hardly worth discussing, because we do know that light rays are curved in a gravitational field, and there is no justification for trying to treat light and material particles

independently. The properties of distance, as exact relations, have already been seen to need some modification when there is relative motion in the system. For an observer may find by trial with measuring rods that two distances along his x and y axes are equal; but to an observer moving along the first observer's x axis these distances will appear different. The concept of distance has a definite meaning only in the absence of relative motion. But ds retains a definite value even when there is relative motion. What we are still entitled to say, then, is that with reference to any observer the position of a particle at any instant can always be specified by three variables x_1, x_2, x_3, the instant itself being specified by a fourth time-like variable x_4. Then we may say that an *event* is specified by the four variables x_1, x_2, x_3, x_4. If two events happen at neighbouring places at a short interval of time, we can say that ds^2 is a quadratic function of the changes of the four variables, the coefficients being functions of the variables. We write then

$$ds^2 = g_{11}dx_1^2 + g_{22}dx_2^2 + g_{33}dx_3^2 + g_{44}dx_4^2 + 2g_{12}dx_1dx_2 + \ldots + 2g_{34}dx_3dx_4 \quad (1)$$

$$= g_{ij}dx_i dx_j \qquad (i,j = 1, 2, 3, 4), \qquad (2)$$

where the g's are to be determined. In (2) we use the summation convention of tensor calculus, that where a suffix such as i or j is repeated it is to be given all its possible values in turn and the results added up. By symmetry we can take

$$g_{ij} = g_{ji}. \qquad (3)$$

In the absence of gravitation we can take $\iota x_1, \iota x_2, \iota x_3$ to be the Cartesian co-ordinates, and x_4 to be ct. Then

$$g_{11} = g_{22} = g_{33} = g_{44} = 1, \qquad (4)$$

$$g_{12} = g_{13} = \ldots = g_{34} = 0. \qquad (5)$$

In presence of gravitation the g's will be modified. We have one consideration from Newtonian dynamics to guide us. The departure of the velocity of a particle from constancy depends on the first space derivatives of the gravitation

potential V. Far away from matter, these are zero, and the particle moves in a straight line. If they are constant, all particles have the same acceleration. Near other matter, these derivatives are not zero or constant; but a function formed from their variations from place to place, namely

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2}, \qquad (6)$$

is still zero outside matter, but finite inside it. Now it appears from 9·4 (6) that the variable parts of the g_{ij} are likely to reduce approximately to multiples of V. Also if the g_{ij} were constants the relation $\delta \int ds = 0$ would imply uniform velocity. Thus the existence of accelerations still depends on variations of the g_{ij}, as before on variations of V; and we look for a set of second order differential equations that may hold outside matter, corresponding to (6).

Einstein's procedure is to notice that the condition that a particle shall move with uniform velocity is equivalent to the condition that ds^2 can, by a transformation of co-ordinates, be put in form such that (4) and (5) hold. With some forms of the original g_{ij} this is possible; with others it is not. When it is possible the field is called Galilean, and the special theory of relativity applies. The condition that it may be possible is that a certain fourth order tensor $B^k{}_{ijl}$ depending on the g_{ij} and their first and second derivatives with regard to the co-ordinates shall be zero. This on the face of it has 256 components, but on account of various symmetry relations only 20 are actually independent. The vanishing of all components of this tensor is the condition for the absence of a gravitational field. Einstein looks then for a set of equations formed from them that may persist in the neighbourhood of matter, but outside it, and a suitable set is found to be

$$G_{ij} = B^l{}_{ijl} = 0, \qquad (7)$$

where in accordance with the summation convention l is given all the values 1, 2, 3, 4 and the results added. Then G_{ij} is a tensor of the second order, with the same number of

components as the original g_{ij}, and its vanishing gives the requisite number of differential equations for the latter. It can be shown that if G_{ij} vanishes in one set of co-ordinates it does so in all, so that we can write down these equations in any co-ordinates we may choose. Inside matter G_{ij} is not zero. When the field is not varying with the time, that is, if all derivatives of the g_{ij} with regard to x_4 are zero, $G_{44} = 0$ is found to be equivalent to $\nabla^2 g_{44} = 0$, apart from a term involving c^{-2}. Within matter, by analogy with Newton's law, we may therefore say that G_{44}, like $\nabla^2 V$, is proportional to the density. The three components G_{14}, G_{24}, G_{34} are related to the momentum per unit volume, and the six G_{11}, G_{12}, ... G_{33} to the six components of stress that occur in the theory of elasticity.

The solution of the equations has actually been carried out completely in only one case, that where the field is symmetrical. In the case of the sun, for instance, we may imagine the time to be that of an observer on the sun, and the direction of a particle specified by the usual angular co-ordinates θ and ϕ. Another co-ordinate is needed to give the distance from the sun. Now if we imagine a short rod placed at right angles to the radius from the sun, it subtends a small angle, $d\psi$ say, at the sun. Its length being $d\sigma$, we say that the distance r is to be given as $d\sigma/d\psi$. Then for such small displacements as make dr and dt zero, we define r by

$$ds^2 = -r^2 (d\theta^2 + \sin^2\theta\, d\phi^2), \qquad (8)$$

and in general

$$ds^2 = g_{11}(r)\, dr^2 - r^2 (d\theta^2 + \sin^2\theta\, d\phi^2) + g_{44}(r)\, dt^2, \qquad (9)$$

for by symmetry g_{11} and g_{44} must be functions of r only. Einstein proceeds to obtain the G_{ij}, and finds that they can vanish only if

$$g_{11} = -g_1(r) = -(1 - 2fm/c^2 r)^{-1};$$
$$g_{44} = c^2 g_4(r) = c^2 (1 - 2fm/c^2 r), \qquad (10)$$

where m is seen, on comparison with Newton's theory, to be identical with the mass of the sun.

It is found that with this form of ds^2 the paths of the planets still agree with those found from the Newtonian law within the errors of observation, with one exception. The new law is found to imply that the path of a planet is not exactly an ellipse, but a slowly revolving ellipse, the direction of the major axis turning round at a constant rate. This change is inappreciable by observation except for the planet Mercury, which was known to have an outstanding departure from the Newtonian theory of just this character; and the amount found from Einstein's theory agreed closely with that already known to exist.

The form of a ray of light near the sun was found to be curved, so that stars would not be seen in quite their usual directions if the light from them to the earth passed near the sun on the way. The amount of the deflexion was calculated, and the amount observed at the total eclipse of the sun in 1919 and at several later eclipses agrees with it.

9·42. The theory is therefore well supported by observation, and the general principle that the paths of particles and light are determined by the behaviour of ds, subject to the coefficients satisfying relations of the form $G_{ij} = 0$ outside matter, is in a strong position. But the other point of view is not exhausted. It can be asked, and often still is, whether any other law than Einstein's will account for the perihelion shift of Mercury and the displacement of star images. This question is habitually ignored in ordinary expositions of the theory of relativity, but it is of capital importance. It is well known that there is matter within the orbit of Mercury, some forming the solar corona and some reflecting the zodiacal light. Such matter is qualitatively capable of accounting for the perihelion movement of Mercury by its attraction, and for the displacement of star images by its refraction. Indeed before Einstein's theory theories were in existence that appeared to account for the anomaly in the movement of Mercury by the attraction of the zodiacal matter, and also for

an anomaly that exists in the motion of the plane of the orbit of Venus*. The latter is of course not touched by Einstein's theory, but it is not very much larger than the probable error, and might just possibly be due to error of observation. If then matter existed in such quantity as to account for any important fraction of the anomaly in the motion of Mercury or of the displacement of star images, the remainder would not be in accordance with Einstein's theory, which would therefore be false. But its amount can actually be estimated from the amount of light that it reflects, and it can be shown† to be much too small to account for any appreciable fraction of the observed effects. These must therefore be due to a departure of the law of gravitation from that of Newton.

The next question is whether, given that the excess motion of the perihelion of Mercury and the displacement of star images are of gravitational origin, any other law than Einstein's would account for them. An answer to this question also can be given. If we return to 9·41 (10) and assume $g_1(r)$ and $g_4(r)$ expanded in series of powers of $1/r$, thus:

$$g_1(r) = 1 + A_1 r^{-1} + B_1 r^{-2} + \ldots, \qquad (1)$$

$$g_4(r) = 1 + A_4 r^{-1} + B_4 r^{-2} + \ldots, \qquad (2)$$

then the equation $\delta \int ds = 0$ is equivalent to

$$\delta \int \frac{ds}{dt}\, dt = \delta \int L dt = 0, \qquad (3)$$

where $$L^2 = -g_1(r)\dot{r}^2 - r^2(\dot{\theta}^2 + \sin^2\theta\dot{\phi}^2) + c^2 g_4(r). \qquad (4)$$

This leads by the methods of the calculus of variations to

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}}\right) - \frac{\partial L}{\partial r} = 0, \qquad (5)$$

and two similar equations in θ and ϕ. We see easily that the θ equation is satisfied if $\theta = \frac{1}{2}\pi$ permanently. The ϕ equation has a first integral

$$\frac{r^2 \sin^2\theta\dot{\phi}}{L} = \text{constant}. \qquad (6)$$

* Jeffreys, *M.N.R.A.S.* **77**, 1916, 112–118.
† *Ibid.* **80**, 1919, 138–154.

There is another first integral,

$$\dot{r}\frac{\partial L}{\partial \dot{r}} + \dot{\theta}\frac{\partial L}{\partial \dot{\theta}} + \dot{\phi}\frac{\partial L}{\partial \dot{\phi}} - L = \text{constant}, \tag{7}$$

which leads to

$$g_4/L = \text{constant}. \tag{8}$$

Then (6) and (8) together, with $\sin\theta = 1$, give

$$r^2\dot{\phi}/g_4 = \text{constant} = h. \tag{9}$$

We can use this to eliminate the time from (8); we get on putting

$$r = 1/u, \tag{10}$$

$$g_1\left(\frac{du}{d\phi}\right)^2 + u^2 = \frac{A}{h^2} + \frac{c^2}{h^2 g_4}, \tag{11}$$

where A is another constant.

This is equivalent to

$$g_1\frac{d^2u}{d\phi^2} + u + \tfrac{1}{2}\frac{dg_1}{du}\left(\frac{du}{d\phi}\right)^2 = \tfrac{1}{2}\frac{c^2}{h^2}\frac{d}{du}\left(\frac{1}{g_4}\right). \tag{12}$$

In the actual motion of a planet u is nearly constant. We put

$$lu = 1 + \xi, \tag{13}$$

where ξ is small and has mean value zero. Then (12) gives to the first order in ξ

$$1 = \tfrac{1}{2}\frac{lc^2}{h^2}\left[\frac{d}{du}\left(\frac{1}{g_4}\right)\right]_{u=1/l}, \tag{14}$$

$$g_1(1/l)\frac{d^2\xi}{d\phi^2} + \xi = \tfrac{1}{2}\frac{lc^2}{h^2}\left[\frac{d^2}{du^2}\left(\frac{1}{g_4}\right)\right]_{u=1/l}\frac{\xi}{l}. \tag{15}$$

Now for planets more distant than Mercury h^2/l is always the same, and can be denoted by the fm of Newton's theory. Thus $\frac{d}{du}\left(\frac{1}{g_4}\right)$ is sensibly constant for l greater than the mean distance of Mercury. Thus it is equivalent to its first term, $-A_4$. Therefore

$$A_4 = -2h^2/lc^2 = -2fm/c^2. \tag{16}$$

Then dividing (15) by (14) we have

$$g_1 (1/l) \frac{d^2\xi}{d\phi^2} + \xi = \frac{\xi}{l} \left[\frac{\frac{d^2}{du^2}\left(\frac{1}{g_4}\right)}{\frac{d}{du}\left(\frac{1}{g_4}\right)} \right]_{u=1/l}, \qquad (17)$$

that is,

$$\left(1 + \frac{A_1}{l} + \frac{B_1}{l^2} + \ldots\right) \frac{d^2\xi}{d\phi^2} + \xi \left[1 - \frac{2(A_4^2 - B_4)}{A_4 l} + O\left(\frac{1}{l^2}\right)\right] = 0. \qquad (18)$$

For planets more distant than Mercury ξ is of the form $e \cos(\phi - \alpha)$, where e and α are constant for each planet. If in general ξ is of the form $e \cos(p\phi - \alpha)$ we have

$$\left(1 + \frac{A_1}{l} + \frac{B_1}{l^2} + \ldots\right) p^2 = 1 - \frac{2(A_4^2 - B_4)}{A_4 l} + O\left(\frac{1}{l^2}\right). \qquad (19)$$

For l large $p^2 = 1$, as it should be. For Mercury,

$$p^2 = 1 - 6fm/c^2 l, \qquad (20)$$

by observation. Substituting for p^2 and A_4 in (19) and equating coefficients of $1/l$ we have

$$A_1 - \frac{2B_4}{A_4} = \frac{2fm}{c^2}. \qquad (21)$$

Now consider a light wave coming from an infinite distance. Then $L = 0$, since $ds = 0$ for two neighbouring positions of a light wave, and therefore in (11), $A = 0$. Also if at a great distance the velocity is c along a line passing at distance a from the centre of the sun,

$$h = -\left[r^2\dot{\phi}\right]_{r=\infty} = -ac. \qquad (22)$$

Thus

$$g_1 \left(\frac{du}{d\phi}\right)^2 + u^2 = \frac{1}{a^2 g_4}. \qquad (23)$$

If $\phi = 0$ when $u = 0$ ($r = \infty$) and we neglect the differences of g_1 and g_4 from unity, a solution is

$$au = \sin\phi. \qquad (24)$$

Then as r decreases from ∞ to a and increases to infinity again ϕ increases steadily from 0 to π.

According to (23)

$$[\phi] = 2\int_0^{au=\sqrt{1/g_4}} \frac{(g_1 g_4)^{\frac{1}{2}} a\,du}{(1 - g_4 a^2 u^2)^{\frac{1}{2}}}, \tag{25}$$

since u has to increase to its maximum and decrease to zero again. Put

$$au\, g_4^{\frac{1}{2}} = x. \tag{26}$$

Then

$$[\phi] = 2\int_0^1 \left\{1 + \tfrac{1}{2}\frac{A_1 - A_4}{a}x + O\left(\frac{A_1^2}{a^2}\right)\right\} \frac{dx}{\sqrt{1 - x^2}} \tag{27}$$

$$= \pi + \frac{A_1 - A_4}{a} + O\left(\frac{A_1^2}{a^2}\right). \tag{28}$$

Thus ϕ increases by more than π during the passage; the ray has a curvature towards the sun. The observed deflexion is $4fm/c^2a$, whence

$$A_1 - A_4 = 4fm/c^2, \tag{29}$$

and therefore

$$A_1 = 2fm/c^2. \tag{30}$$

From (21) now

$$B_4 = 0. \tag{31}$$

It follows that

$$ds^2 = c^2\left\{1 - \frac{2fm}{c^2r} + O\left(\frac{2fm}{c^2r}\right)^3\right\} dt^2 - \left\{1 + \frac{2fm}{c^2r} + O\left(\frac{2fm}{c^2r}\right)^2\right\} dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{32}$$

Einstein's solution was

$$ds^2 = c^2\left(1 - \frac{2fm}{c^2r}\right) dt^2 - \left(1 - \frac{2fm}{c^2r}\right)^{-1} dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{33}$$

so that all the terms in it capable of producing a perceptible effect are directly demonstrated by the observational data.

9·5. It might appear that as Einstein's law of gravitation was obtained as a result of his considerations of the general relativity of the laws of nature, and afterwards verified by

observation, the last discussion is of the nature of a prophecy after the event. I think, however, that it was really rather in the nature of an accident that Einstein's law was obtained by his method and not by one very like that just given. The motion of the perihelion of Mercury had been known since Leverrier's theory of the planetary motions, and it was known that a slight modification of Newton's law of gravitation would account for it. The only one actually suggested was that of Asaph Hall, in which the index in the law was made slightly different from 2. If the simplicity postulate had been explicitly stated at that time it would have been recognized that a law with an index slightly different from an integer is in reality an extraordinarily complicated one, and has therefore so small a prior probability that it does not merit serious consideration. The alternative that should have been tried was to include in the gravitational force terms varying inversely as the third and fourth powers of the distance, and to choose the coefficients so as to account for the facts. It would then have been found that the ratio of the coefficient of the third power to that of the second was of the order of fm/c^2; and a direct relation between gravitation and light would have been indicated. Such a relation had been tentatively suggested by Newton himself and by Laplace. The curvature of light rays passing near the sun had indeed been predicted by Newton. His suggestion had been forgotten; but a discovery of this sort would certainly have revived interest in it and led to an experimental test. It would then have been found that the deflexion was twice what he predicted, and it would have been seen that a more drastic revision of Newton's law was necessary than the mere addition of a cube term to the gravitational acceleration. In that case every essential of Einstein's law would have been obtained before his theory was created, and his result would have been merely a mathematical description of facts already known. Of course the fact that Einstein was able to construct his theory without such previous considerations is an additional reason for admiring Einstein.

It might be said that in inferring the law of gravitation from the empirical facts we have gratuitously assumed that the departures from the Newtonian law arise from the terms of the lowest orders in the g's that are not considered in the first approximation. But if they arose from later terms extraordinarily large numerical coefficients would be needed, which again are excluded by the simplicity postulate.

9·6. There is no antagonism between the principle of relativity and the simplicity postulate; indeed the principle is itself a thoroughgoing application of the postulate. The simplest possible relation between two quantities is that they are independent; that is, that when one changes there is no associated change in the other. When there is an associated change we may either study it directly, or try to construct a new quantity that does not change. In Newtonian dynamics, for instance, the velocities of the bodies in a system change with time. We can proceed to find new quantities, the momenta of the system, which do not change with time. We may deal with the kinetic energy either by saying that

$$\text{change of kinetic energy} = \text{work done},$$

or we can reverse the sign of the work and introduce an apparently new concept, the potential energy, and say that

$$\text{kinetic energy} + \text{potential energy} = \text{constant}.$$

Actually this procedure is of doubtful legitimacy, because the work done may depend on the mode of passage from the initial to the final state, in the case of non-conservative forces, and then the existence of potential energy as a function of the state of the system is problematical. It is found, however, that the operation of non-conservative forces involves the generation of heat, and that there is a relation between the work done by these forces and the amount of heat produced. A new kind of energy, heat energy, is then invented, and we say that

$$\text{kinetic energy} + \text{potential energy of conservative forces} + \text{heat energy} = \text{constant}.$$

So we proceed, inventing new kinds of energy so as to keep the principle of conservation of energy true. It is actually found at each stage that the new kinds of energy have definite properties of their own that warrant their being regarded as physical concepts. In modern molecular physics heat energy itself has come to be regarded as kinetic and potential energy of agitation of the molecules, thus making it possible to regard all forces in the last resort as conservative; the distinction is then between large-scale or molar motion that we can observe in dynamical experiments, and small-scale or molecular motion that is not directly observable as motion, but can be detected either by the sensation of heat or by the production of thermal expansion. In all stages we detect the operation of a prior probability that *something* is constant, and that our problem is to find out what it is.

But the conservation of energy and momentum does not comprise the whole of dynamics. To find the actual relative motion of the parts of a system we still need the equations of motion, or their equivalent, Hamilton's principle; and these imply the conservation of energy and momentum but are not implied by them. We prefer Hamilton's principle to the equations of motion because it is expressible directly in terms of the ultimate ideas of distance and time, whereas the equations of motion, apparently at least, involve the co-ordinate system. Hamilton's principle is not a conservation principle; the essence of it is that the integral involved in it is *not* stationary if the path we begin with is anything but a dynamically possible path. But we do notice in it that the statement is independent of the system of co-ordinates; and it is this fact, applicable to the whole well-verified region of Newtonian dynamics, that is the basis of the belief that the ultimate laws of nature can be stated in a form independent of the co-ordinate system. By the simplicity postulate we are therefore entitled to say with a high probability that this principle is true in general. But this principle is precisely the principle of general relativity. The special theory of relativity

enables us to extend this to the motion of light, without disturbing it for material particles outside of a gravitational field, only an ultimate physical constant, the free velocity of light, appearing. But in the special theory it turns out that dynamical time is on a very similar footing to the three position co-ordinates, and the principle of relativity has to be extended to allow the system of reference to include *four* variables, three position co-ordinates and the time, which can then be transformed freely for systems of reference in motion relative to each other. Incidentally it turns out that there are not two fundamental ideas, distance and time, but one fundamental idea, ds; and that the coefficients in ds^2 in the general theory satisfy differential equations that have the same form in any transformation, just as the gravitation potential of Newtonian dynamics satisfies a differential equation that is not affected by changes of axes.

It must be said that in spite of the high probability we can now attach to the principle of general relativity, it is still on its trial. It is really verified so far only in the motion of a body or light ray of negligible mass in a symmetrical field. The obvious generalization to a system of many particles would be to attribute a mass and a ds to each, and to say that $\Sigma \int m_i ds_i$ is stationary. But a deeper analysis appears to be necessary, and the application of the principle to even the problem of two bodies has not yet been carried out, on account of the mathematical difficulties.

A further problem concerns the size of the universe. The coefficient g_{44} outside a spherical body, we have seen, is $c^2(1 - 2fm/c^2r)$. If the universe has density ρ and radius a, then just outside it g_{44} will be $c^2\,(1 - \frac{4}{3}\pi f\rho a^2/c^2)$. This is negative if ρa^2 exceeds a certain value, and the corresponding local velocity of light is imaginary. There is thus a definite upper limit to the size that the universe can have if its mean density is given; and there is a lower limit to its size if its total mass is given. Various solutions of the problem have been attempted, but there is so far no definite answer.

The chief outstanding problem, however, is in relation to electricity and magnetism. It was shown by Lorentz and Larmor that the equations satisfied by the electric and magnetic forces satisfy the special theory of relativity, and numerous electrical experiments designed to detect absolute velocity led to null results. But light waves are electromagnetic in character, and are now known to be affected by gravitation. Thus gravitation and electromagnetic phenomena interact, and the question is, do the laws of this interaction satisfy the general theory of relativity? The question really presupposes a condition analogous to the classical "First catch your hare". We cannot test these laws experimentally until they have been produced, and although several experts have produced theories they do not seem to have satisfied one another.

The general theory of relativity is therefore justified as a physical law up to a certain point, and the simplicity postulate entitles us to extend it further, if possible. This extension is a matter for the future and for further experimental investigation.

CHAPTER X

MISCELLANEOUS QUESTIONS

"How is bread made?"
"I know *that*!" Alice cried eagerly. "You take some flour——"
"Where do you pick the flower?" the White Queen asked. "In a garden, or in the hedges?"
"Well, it isn't *picked* at all", Alice explained; "it's *ground*——"
"How many acres of ground?" said the White Queen. "You mustn't leave out so many things."

LEWIS CARROLL, *Through the Looking Glass*

This chapter is devoted to a number of incidental considerations that have so far escaped attention.

10·1. *Is there a non-quantitative simplicity postulate?* Let us consider such a biological proposition as the following. *All animals with feathers have beaks, two legs, two wings, and warm blood.*

We might try to analyse this as a proposition in sampling. "Being an animal with feathers" would then be the property a of a class, m members of which have been observed. "Having a beak" is taken as the property b possessed by l of the observed members. Then according to Laplace's theory of sampling the probability that the next member examined has the property b is $\frac{l+1}{m+2}$; and if all the observed members have had the property, so that $l = m$, the probability that the whole class, of number n say, has the property b is $\frac{m+1}{n+1}$.

This result, in relation to the proposition under consideration, seems to be at variance with general opinion. I have observed a large number of animals with feathers, but I suppose that they constitute less than 1 in 10,000 of the animals with feathers in England. According to Laplace's theory, then, $(m+1)/(n+1)$ is under $\frac{1}{10000}$, and the prob-

ability that I should attribute, on the data, to the proposition that *all* such animals in England have beaks does not exceed this trivial fraction. Actually I seem to attribute to it a probability approaching certainty. The same is, I believe, the position of most ornithologists. Our problem is to consider the reason for this great departure from Laplace's theory.

It might be thought that since "animal with feathers" is so widely recognized a concept as to have had the name *bird* associated with it, and as such to have been mentioned in literature on countless occasions, the information provided by other people contributes largely to one's estimate of the probability. If such a concept is generally recognized, a "bird" without a beak would attract attention and be commented on, and the absence of comment gives some ground for supposing that nobody has seen such a thing. In this particular proposition such considerations certainly carry weight; but I do not consider them the ultimate reason. In the first place, other people's judgments are not known to me directly; the things that I know directly about other people are their appearance and the sensations of sound that they produce, and the appearance of the marks they make on paper. When I attribute to their sounds and writings meanings similar to those I express when I make similar sounds and marks myself, I am making an inference. It seems to me that such an inference must rest again on observed similarities between other people's behaviour and my own, which are generalized as part of the science of psychology, and depend for their acceptance as general propositions on a theory of sampling. I have not examined a large fraction of the inhabitants of England to find out whether they do seem to attribute the same meanings to propositions that I do, and when I assume that those I have not examined do so I am making an inference, which on Laplace's theory would itself have as small a probability as the proposition about birds. The introduction of testimony therefore only shifts the issue without affecting its ultimate nature.

The situation occurs, indeed, when no question of testimony arises. A botanist finds a plant that does not fit any description already recorded. He immediately calls it a *new species*, and publishes a description of it; that is, a new concept is created on the basis of a single observed instance. There is no question of anybody else having observed the same species. But we notice that one property does not make a species. If a botanist found, for instance, a plant agreeing in all particulars with the descriptions of the upright buttercup, but possessing no petals, he would not publish a description of a new species, *Ranunculus apetalus*. He would call it a specimen of *Ranunculus acer* without petals*. The mere possession of one unusual property does not constitute a new species, but merely a freak. There must be a conjunction of several new properties, and then it is expected that some at any rate of these will always be associated in future instances. It is utility for purposes of prediction here, as in quantitative laws, that coincides with the introduction of a new concept. The principle seems to be that if an object of a given class has r properties $a, b, c, \ldots k$, then there is a finite prior probability that all future members of the class with any $r - 1$ of these properties will also have the remaining one. This probability is a moderate number, independent of n the number of members of the class with $r - 1$ of the properties in the world. If it was merely $1/n$, we should be back to Laplace's theory; and we seem to have reached again the principle that Laplace's assessment of the prior probability is wrong for the extreme cases where all or none of the members in the world have the property under discussion. But if the prior probability is moderate, say $\frac{1}{3}$, whatever n is, it appears that repeated verifications will make the probability of the law approach certainty, as for quantitative laws. Then we have a simplicity postulate applicable to non-quantitative laws.

* Unless he forgot his *acer, acris, acre* and called it *Ranunculus acris*.

We may hazard a solution of this question by considering the prior probability that a point may lie within a given interval on a line. If the line is infinite in length both ways, so that there is nothing to distinguish one interval from another, the prior probability is uniformly distributed and the probability that the point lies in a given interval is proportional to the length of the interval. Strictly this makes the probability that it may lie in any finite interval infinitesimal, but we need consider only the ratios of the probabilities for different intervals, which are perfectly definite; and when measures are introduced factors such as $e^{-h^2x^2}$ enter into the inverse probability and make the posterior probabilities definite. If the point is restricted only to lie to the right of a given origin on the line, and we have no previous knowledge about its distance x from that origin, the prior probability that it lies in a short interval dx is proportional to dx/x; for with any other law the probability that it lies between x and $2x$ would depend on x, and therefore there would be a previous criterion suggesting a scale of distance. Now suppose that we know initially that x lies between 0 and 1. The prior probability that it lies in a range dx must be symmetrically distributed about the centre of the range, so that it must be of the form $f\{x(1-x)\}\,dx$. But when x is small the influence of variable distance from 1 must be inappreciable, and therefore when x is small

$$f\{x(1-x)\} \propto 1/x.$$

But in this region $f\{x(1-x)\}$ is nearly $f(x)$, and the required law is therefore that the prior probability that x is in a range dx is proportional to $dx/x(1-x)$; and integrating this we see that the prior probability that it lies between a and b $(b > a)$ is proportional to $\log \frac{b}{1-b} - \log \frac{a}{1-a}$. The fact that this tends to infinity when b tends to 1 or a to 0 is not really serious, because in actual measures the extreme values are usually excluded by the limits of error. Now this suggests a

form of the prior probability in the theory of sampling. We are given there that a ratio $x = r/n$ is at least 0 and at most 1. Then the prior probability of a given value of r may be taken as proportional to

$$\frac{1}{x(1-x)} = \frac{n^2}{r(n-r)}.$$

Thus the $f(r)$ of the theory of sampling should be taken to be inversely proportional to $r(n-r)$. This makes $f(r)$ tend to infinity at the extreme values; but as before this is not serious, for so long as the sample is homogeneous the extreme values are still admissible, and we *do* attach a high probability to the proposition that the whole class is of one type; while as soon as any exceptions are known the extreme values are completely excluded and no infinity arises. Such a form of $f(r)$ seems therefore to be just what is needed to provide a simplicity postulate for non-quantitative laws.

It may happen that an observed new conjunction of properties breaks down in further instances. This is precisely the case where the botanist cannot find permanently associated characters to describe his species properly, and occurs in such "difficult" genera as *Rosa* and *Hieracium*. From our point of view these are instances of suggested laws, with finite prior probabilities, that have broken down under crucial tests.

It appears that such a principle is of great importance in the theory of our knowledge of the world, and that the validity of even the concept of *objects* itself depends on it.

If we return to the notion of a bird now, we see that *feather* really expresses in itself the conjunction of numerous properties. A feather has a central horny quill, fringed by numerous filamentous hairs so arranged as to lie side by side nearly in a plane, and so that their ends lie on a smooth curve. It is this conjunction of properties that justifies the introduction of the concept and the attachment of a definite name to it. Similarly *beak* implies a horny projection on the face, carrying with it a mouth and nostrils; again several properties are associated. The observed usual conjunction of

the two sets of properties, and also those of two legs, two wings, and warm blood, warrants the inference that the conjunction holds in general and therefore the introduction of the concept *bird*.

It happens that, so far as our knowledge goes, all animals with feathers have the other properties mentioned; the converse is not true. Thus the duck-billed platypus has a beak and warm blood, but has not feathers and has four legs and no wings; man has two legs and warm blood, but not a beak or feathers. What if there were no single defining property; if, that is, there were animals with a beak, two legs, two wings, and warm blood, but covered with hair instead of feathers? Should we then have to abandon the notion of *bird*? I think not. We should call the new creatures *birds with hair*, just as we call the duck-billed platypus a *mammal with a beak*; or else we should retain the definition in terms of feathers and deny that the new creatures are birds at all. There would probably be a vigorous discussion in the zoological journals as to which course was the correct one, but in any case the decision is a matter of convention, like the assignment of a name to the concept in the first instance. The important thing is the observed usual conjunction of the properties, upon which we base the inference that the properties are likely to be associated in future instances. The existence of an occasional exception does not disprove the rule; it merely suggests new lines of inquiry. The concept is merely a way of expressing the rule concisely.

After the above passage had been written I came upon the following, in a paper by Dr A. Wohlgemuth*.

"The point has been admirably stated by Freud's colleague, Joseph Breuer:

"'All too easily one gets into the habit of thought of assuming behind a substantive a substance, of gradually understanding by consciousness an entity. If then one has got used to employing local relations metaphorically as, e.g., *subcon-*

* *J. Medical Psychology*, 5, 1925, 105.

scious, as time goes on an idea will gradually develop in which the metaphor has been forgotten, and which is as easily manipulated as a material thing. Then mythology is complete.'

"Breuer recognized the slippery slope down which Freud rushed away from scientific fact, and called a warning halt, but, alas, too late."

With the statement of psychological fact that we do get into the habit of assuming a substance, or, as I prefer to say, a concept, behind observed conjunctions of properties, and that the concept comes to be manipulated as directly and easily as a material thing, I am in complete agreement. From the statements of opinion by the two authors quoted that this occurs *too* easily and that it constitutes a means of rushing away from scientific fact, I dissent completely. These opinions are a direct negation of the whole of the scientific procedure of constructing concepts, and there would be no such thing as science, or, indeed, as everyday knowledge, if they were accepted. It is precisely the utility of concepts in summarizing existing knowledge that makes it possible to keep scientific facts classified and accessible, and therefore to make progress as new laws are discovered. The existence of dynamics, for instance, depends first on abstracting the concepts of physical objects and events from sensations; then on the concepts of intervals of time and distance, derived from events and objects; then on those of mass and force, derived from the observed relations between distances at different times. At each stage the concept gets further away from the original facts; but at each stage also it makes it possible to infer more facts. The double aspect of the construction of concepts is not antagonistic to scientific method, but on the contrary is the very essence of it.

10·2. *Ultimate concepts.* In the development of knowledge our fundamental data are sensations and certain *a priori* principles of logic and probability, and as we proceed we construct

concepts of increasing generality from them. Is there any reason to suppose that the process will ever stop? If so, the concepts reached at this final stage may be called ultimate concepts. It has happened that for ages certain concepts have been thought ultimate, but are now proving to be expressible in terms of more general ones. Thus distance and time, which were long thought to be ultimate and absolutely general, are found to be approximations involving a certain amount of ambiguity, the more general concept behind them being the *ds* of the theory of relativity. The physical object itself, with its characteristic dynamical property of impenetrability, is no longer a continuous piece of "matter" occupying a definite region of space. It lost that status when Dalton showed that the simple numerical relations that arise in the laws of chemical combination could be explained if matter consisted of molecules, each molecule of a definite substance consisting of a finite number of atoms, the total number of kinds of atoms being finite. It followed at once that a piece of a chemical compound could not be, in the last resort, a region of space with the same properties at all points. The molecular theory led further, in the hands of Waterston, Boltzmann, Maxwell, and others, to mechanical explanations of Boyle's and Charles's laws in gases, and the viscosity, diffusion, and thermal conductivity of gases. In modern physics, experiment in rarefied gases reaches directly not only the molecule, but the atom; and even the atom proves to have properties explicable on the hypothesis that it is made up of only two kinds of entity, the proton, positive nucleus or hydrogen ion, carrying a positive charge, and the electron, carrying an equal negative charge. Application of this notion of the ultimate constitution of matter to solid crystals has led W. H. and W. L. Bragg to explanations of their behaviour in reflecting X-rays, and Max Born and J. E. Lennard-Jones to quantitative explanations of their elastic, optical, and electrical properties. The principle that matter is made up of protons and electrons is therefore in a strong position. But the impenetrability of

matter has lost its generality. In a gas under ordinary conditions the region actually occupied by the molecules is under a thousandth of that of the whole; even in a solid the protons and electrons do not occupy more than an exiguous fraction of the whole region within the apparent outer surface. We cannot as a matter of fact push one piece of matter through another without meeting a resistance; but this resistance is explained by the theory. Further, electrons can be made to pass right through films or plates of solids, finding their way between the constituent protons and electrons.

These modern views on the constitution of matter did not lead directly to the abandonment of the idea of a physical object as an ultimate reality, but rather to the attitude that the object, as usually understood, is composed of smaller things, which are still objects; that is, like the physical object before the time of Dalton, they have definite positions at any time, and no two of them can occupy the same region. But even this position is being assailed by the new quantum mechanics. According to Heisenberg's uncertainty principle, which is a simple consequence of any quantum theory, it is never possible to measure the position and velocity of a particle accurately at the same time; whichever of them we try to measure, the process affects the other, and an indeterminacy remains in both. Relativity has left us thinking that an event can be specified by stating exact values of four variables, three position co-ordinates and the time. Heisenberg leaves us in doubt as to whether these variables can have any exact values at all; and if the position of a particle is indefinite it becomes doubtful whether the statement that two particles cannot be in the same position has any meaning.

In the various forms of the new quantum mechanics the four variables needed to specify the time and the position of any particle have ceased to be physical magnitudes at all; a single numerical measure is not enough to specify any one of them. In Heisenberg's theory each is replaced by a matrix, an assemblage of several magnitudes; in that of Dirac the

co-ordinates and the corresponding momenta are what he calls q-numbers, which do not satisfy the ordinary rule of multiplication $pq = qp$. In the theory of Schrödinger an entirely new variable, the wave-function ψ, appears, which satisfies a certain differential equation, and the observed phenomena of electron emission, radiation, and so on emerge as expressions of the properties of the wave function. All three theories appear to give the same answers, and to be well confirmed by experiment. But all agree in that the ultimate particles do not have definite co-ordinates at any instant. The proton and the electron, as particles with definite positions, have disappeared. Whereas on the older quantum theory a hydrogen atom consisted of one proton with one electron moving in a definite orbit about it, on the new theories the proton and electron have lost their individuality and can only be said each to fill the whole region occupied by the atom. As we cannot observe the positions of the electron at various points of its alleged orbit, and should certainly alter the orbit if we tried, there is no experimental objection to this view. It is only when the electron emerges from an atom and travels freely that it behaves as an individual, and in these conditions the new theories describe its actual behaviour.

On Schrödinger's theory the co-ordinates appear explicitly in the differential equation, though the electron as a thing with definite co-ordinates has disappeared. Thus the notion of position in space remains though nothing has a definite position. This situation is somewhat paradoxical, and an attempt has been made to overcome it by constructing from ψ a real function, which is said to represent the probability that a given position is occupied at a given time. Thus we have to speak in our ordinary sense of the probability of Schrödinger's differential equation as a scientific law, and yet the equation itself deals with probability. We are in the position of having to speak of the probability of a law of probability. The complication is not really a new one, because

it arises in the treatment of the law of error when the standard error has to be determined from the observations. In that case we have had to speak of the prior probabilities of different standard errors, that is, of different laws of error, where each law is itself a statement about the distribution of probability among different possible values of the error. Perhaps, in addition to the *a priori* laws of probability that underlie all inference, there are other laws of probability that have to be found as far as possible from experience and therefore have probabilities themselves.

The question does not arise in the theories of Heisenberg and Dirac, for the co-ordinates in them are not single real magnitudes. The position has just the same degree of indefiniteness as the particle that is said to occupy it*. This consideration, combined with the formal simplicity of Dirac's theory, seems to place it in the best position of the three. But the formal simplicity of Dirac's laws does not always make it easy to solve his equations in special cases, and it is often found that the solution of his equations is most easily obtained by Schrödinger's method. This is really because Schrödinger's method uses only ordinary mathematics, while Dirac's numbers that do not satisfy the commutative law of multiplication require the construction of a new branch of mathematics, which is not yet fully carried out.

The existence of three such theories, all giving results in agreement with the facts, but formally quite different, leaves us in considerable doubt about ultimate concepts. A fruitful source of philosophical discussion is the reality of scientific concepts. So far as I can see what is usually meant by this is the existence or otherwise of atoms, electrons, light waves, and so on as ultimate realities in the same sense as physical objects appeared to be ultimate realities to the eighteenth-century physicist. The answer to this seems to be definitely in the negative; but the question is replaced by that of the

* The indefiniteness is much like that in the statement that the equation $(x-3)^2+\frac{1}{16}=0$ has roots *near* $x=3$, though actually there is no real root.

reality of co-ordinates, momenta, and wave-functions. It seems to me that this question may well be postponed till we have made more progress with the various new quantum theories, particularly in the direction of co-ordinating them with the general theory of relativity. In any case the concepts that appear explicitly in the theories are quite different in character from physical objects. From the standpoint of scientific method the one and only test of the validity of concepts is whether the laws they are supposed to satisfy explain our sensations; whether this is also a ground for attributing philosophical reality to them is a different question.

10·21. The question of ultimate concepts arises again in such biological questions as the materialistic interpretation of physiology and the physiological interpretation of psychology. Modern research has shown that many physiological processes satisfy quantitative laws like those of physics and chemistry, and in many cases that these processes can actually be interpreted in terms of physics and chemistry. Are we justified in inferring that all physiology is reducible to physics and chemistry? It must be remembered that when the question was formulated the atom was considered an ultimate reality; the result of modern developments in physics is that we are asking whether physiological processes can be explained in terms of q-numbers or ψ-functions. The alternative is that there is a non-physical concept, which we call *life*, and which may be ultimate. The problem of materialism is to explain life. Life as it stands is a valid scientific concept because it explains observed phenomena; a live animal has different properties from a dead one. That is not to say that it is an ultimate concept. There seem to me to be two relevant indications, pointing in opposite directions. The growth of green plants involves the interaction of carbon dioxide and water to produce sugar or starch and oxygen, a reaction requiring the absorption of energy, which the plant obtains from the sun's radiation. Carbon dioxide and water are

ordinarily stable in each other's presence; the plant must apparently have some *directing* ability, applying the solar energy in just such a way as to upset this stability. The same applies to the obscure organisms that derive their energy from chemical reactions without the presence of light, reactions that do not take place spontaneously, but only under the influence of the plant itself. On the other hand, if organisms have a directing power, of molecular fineness, as this would suggest, they might apply it to the sorting out of molecules according to their velocities. Then they could upset the second law of thermodynamics and provide for themselves all the available energy they need. This does not appear to happen; physiological processes in animals and plants appear to follow the second law of thermodynamics. The hypothesis that life is not an ultimate concept remains untested.

10·22. Our primary data being sensations, it may be said that the aim of science is to account for sensations in terms of ultimate concepts and their properties. On the materialist theory these ultimate concepts are those of physics and no others. The physiological interpretation of psychology does not go so far as this, but states that psychological phenomena can or will be reduced to physiology. The experimental study of sensation has gone some way in the explanation of the transmission of sensations to the brain, but little has been done towards understanding what happens to them when they get there. The opinion that the amazing complexity of mental processes, including recognition of sensations, emotions, reasoning, and volition, can be reduced to physiological processes, is hypothesis; it may be true or not, but it is certainly at present pure unverified hypothesis. Further, all the mental processes just mentioned have in common the fact that they involve, to varying degrees, conscious criticism. This is directly recognized and therefore is a fundamental concept. One way of studying it is to examine mental

behaviour when it is removed as far as possible, and to see what differences arise.

It is therefore a legitimate procedure to study mental behaviour when conscious criticism is, as far as possible, eliminated. The absence of criticism is best realized in dreams and in the psychoanalytic situation, where the patient, as a regular matter of technique, says everything that comes into his mind without criticism. The results are not chaotic; they are found to arrange themselves according to perfectly definite rules of resemblance, which are scientific laws. They differ from the rules of conscious criticism, the function of which is to observe and study them; and they are found to be closely related to the forgotten experiences of childhood and the pitiless logic, based on incomplete data, of the *enfant terrible* and the child at still earlier ages. The result is the discovery of a whole region of mental activity, with laws of its own, and demanding new concepts to express them. Freud's Unconscious is the general name for this region; for details of its structure reference must be made to the special literature of the subject*.

The results of psychoanalysis have been criticized on various grounds, which seem to me to merit discussion here because they involve points of principle applicable to any science. One line of attack is simply to deny the facts as discovered, or the truth of the relations found between them. This is merely a matter of refusal to investigate, and does not impress the analyst who is dealing with the material every day, or the patient who has been cured of various mental disorders, ranging from minor anxieties to phobias or disabling neuroses, by being enabled to understand his own mental processes better.

A more subtle attack is to say that psychological processes are really the expressions of physiological ones, and that the solution of the problems investigated must come ultimately from physiology. This may be true. But to use it as a basis

* See especially Freud, *The Ego and the Id.*

of procedure is not legitimate, because it assumes from the start that there are no ultimate mental concepts, or, what is the same thing, it takes for granted that there *are* relations that completely determine the phenomena of conscious mental activity in terms of those of physiology before we know what they are. Instead of inferring the laws from the data, the invariable scientific procedure, it begins with unstated laws and treats the data as a ground for optimism about the future. The situation is the same as if an engineer in process of designing a bridge was told that he should not attend to experimental evidence about the strength of his materials because all phenomena of elastic fatigue, like other elastic phenomena, may some day be explained in terms of modern atomic and quantum theory. It may be so; but he wants to get the bridge built.

It has also been said that the phenomena are not quantitative and therefore not scientific. This consideration would obviously invalidate the greater part of biology; but it would also apparently invalidate the notion of the physical object itself. Quantitative study always rests on a basis of facts recognized qualitatively, and the fact that we cannot as yet measure emotions quantitatively and predict their measures is no ground for saying that emotions do not exist when we know perfectly well that they do, or that they obey no scientific laws when considerable knowledge of those laws has in fact been attained.

A further consideration is that even if such a hypothesis is correct we should still be under an obligation to investigate whether its consequences are true. That implies investigating mental phenomena, and providing explanations of the facts that psychoanalysis has already disclosed. The hypothesis saves no work, but merely attempts to delay it.

10·23. The criterion of philosophical reality has been put in the form "do things exist when they are not observed?" From our point of view this is scarcely a question at all. Our

primary data are sensations, which definitely do not exist when they are not observed, and *a priori* laws, which have the property of truth whether we know them or not. The reality of concepts, on the other hand, is not explicitly involved in the question. To ask whether a physical object exists when it is not observed assumes that it sometimes *is* observed, and this is untrue. The physical object exists only in the sense that it helps to explain sensations; it is never observed directly. To say that "we observe an object" is really a shorthand for saying that we have a series of sensations which are co-ordinated by forming the concept of an object.

In another form, however, the question is significant. We observe the direction of the planet Jupiter at various times and predict its position at other times. We also observe a minor planet and predict its position at any time, allowing for the attraction of Jupiter on it in the meantime. The results are verified irrespective of whether we actually do measure the position of Jupiter in the meantime. It is of course well known that Neptune and the companion of Sirius were discovered through their perturbations of Uranus and Sirius respectively; their gravitational effect was known before they had been seen at all. Our most direct reason for saying that Jupiter or Neptune exists is that we can see it if we take the proper steps; but the motion of other bodies due to it is the same whether we actually observe Jupiter or Neptune in the meantime or not. There is no reason in principle for choosing the direct visual sensation rather than the perturbative effect as our ground for forming the concept of Jupiter or Neptune. The two grounds express co-ordinations of different sets of sensations, that is all. If a concept is formed as a result of one law, and subsequently a second law is found to be true in terms of it, we may often just as well take the second as the definition of the concept; and the two together express a greater generality of the concept than is implied by either alone. In this case, by visual observations, we infer the law of gravitation, giving certain co-ordinates, which we say express

the positions of the planets at any time with a very high degree of probability. These co-ordinates exist at intermediate times because we can calculate them, and when observations are made the inferred values are found correct; no further justification is necessary.

We nevertheless need to allow occasionally for the possibility that certain events may occur only when opportunity arises for observing them. Thus a traveller observing the United States from the train alone might be pardoned for inferring a law* that a bell is always ringing at railway crossings. What he observes is that this law holds when his train is near a railway crossing; he has no opportunity of observing that the bell rings *only* when a train is near. On the face of it this is a case of error introduced by the nature of the observing instrument, but so extreme as to be trivial. Yet it is quite analogous to a stage in the development of modern physics. At the time of the Michelson and Morley experiment physicists generally believed in an all-pervading ether, which transmitted electric waves, including light waves. The experiment, like many others, was designed to detect motion relative to this ether. The failure to obtain any positive result led to the opinion that, though there must *be* an ether, whenever we tried to detect motion with respect to it circumstances conspired to make it impossible to do so. Physicists holding this view were effectively saying that the observer was *always* on the train, however hard he might try to get off it, but were nevertheless clinging to the view that there were times when no train was near and that it was reasonable to speculate about observations in such conditions. Einstein's great advance in 1905 was to recognize from the weight of evidence that a stage had been reached when too many conspiracies of circumstances had to be assumed, and that it was better to take the known facts as they stood and generalize from them.

In the last resort we can never exclude this type of difficulty entirely. The existence of sensations implies the exist-

* As usual, I mean a scientific law, not a legal one.

ence of an observer, and there is therefore always a theoretical possibility that his presence produces effects that do not exist otherwise. The practical reasons for ignoring this possibility are, first, that the presence of another observer does not as a rule alter the observations made by the first, which we should expect to happen if the observer had a disturbing influence; and second, that we do as a matter of fact proceed by describing and inferring sensations, and that the state of the world when not observed is not really relevant. But it is relevant that our laws lead to correct inferences whether or not we have in each experiment checked every intermediate step. When the constant of a tangent galvanometer has been determined in terms of the rate of deposition of copper in an electrolytic cell, it is unnecessary to redetermine it during every later experiment with that galvanometer. The scientific law being once established, subsequent inferences from it are made with the full probability of the law, and repeated verification is not needed.

The influence of the observing conditions is seen again in the difference between experiment and observation. In a laboratory experiment there is usually a possibility of repetition; a result having once arisen, the apparatus can be set up again and we can see whether the same result ensues. If we have a prior belief in determinism we should expect this. Personally I do not think that a belief in determinism is *a priori*. What I think is established by such a repetition is that the result is independent of the time of the experiment. In astronomy, on the other hand, we cannot start the planets off again as they were and see whether they again describe the same orbits. This possibility of control over the initial conditions constitutes the difference between experiment and observation. It is a difference of technique, and not of principle. In the astronomical case it is equally well established that the accelerations are determined by the relative positions of the bodies and do not involve the time explicitly. If we could control the initial conditions, it might have been established with less trouble; but it *is* established.

Such judgments of independence are much commoner in scientific inference than are ordinarily realized. In describing the result of an investigation we tend to restrict our specification to the variables actually found to be relevant. In an electrical experiment we do not usually specify the time of day, the temperature outside the laboratory, the observer's age, or the number of the laboratory assistant's children. The reason for this is not a guess or a prior certainty that these factors are irrelevant. The reason is that in different experiments these factors are actually different, and the results are found to be the same. Now independence is the simplest possible scientific law, corresponding to the simplest differential equation

$$\frac{dy}{dx} = 0.$$

Hence it always has a considerable prior probability, and therefore reaches practical certainty with a very small number of verifications. We expect things to be independent until the contrary is shown; our interest is in discovering relevant variables, not in adding to the enormous number of irrelevant ones.

The belief in determinism is related, I believe, to what philosophers call the Principle of Causality. It may be expressed in the form: given the state of the world at any instant, the state at any subsequent instant is determinate. The truth of the principle requires some discussion of the meaning of *state*. The positions of all particles in the world at some instant obviously do not determine the motion afterwards unless we also know their velocities. But particles are not enough. The *time* must be the time of the event as understood in the theory of relativity. If a light is extinguished at an instant, it is still seen for a finite time at distant places; the sensations produced are the same as if the light was still shining. If we consider the state of the system at an intermediate time, we must say that the illumination seen is caused by the light *on the way*, for the lamp is no longer available as a cause. The state to be

specified as determining the future must therefore include the positions and directions of motion of light waves. The alternative is to say that the state of a system at any instant is determined, not by the state at each single previous instant, but by the aggregate of states at all previous instants. The position is tenable; but now we see that the previous instants to be considered stretch right up to the instant of observation, and we may reasonably say that the state then is determined by the states at intervals indefinitely shortly before. But then the notion of light on the way becomes a necessity, and we may as well say at once that the law of causality is expressed by differential equations with regard to the time. If we insist on specifying the state only in terms of material particles we must consider laws as involving finite intervals of distance and time explicitly, and we meet the ancient question of action at a distance. It seems to me that the answer to this question can be given in terms of the principle that the form of quantitative laws is differential. The form of the properties of light away from a gravitational field is given by Maxwell's differential equations. The fact that light has a constant velocity is a property of the solution of these equations. Thus the fact that two events at the same time at different places do not influence each other is a result explained by the law and scientifically valuable as helping to establish the law. It is at this point, I think, that Robb's theory of conical order of events has its application. Again, the fundamental form of the law of gravitation is

$$\nabla^2 V = -4\pi f\rho$$

on Newton's theory, or its analogue on the general theory of relativity; the integrated form

$$V = \Sigma \frac{fm}{r}$$

is not the fundamental form of the law, but its solution. Action at a distance seems to imply that the latter should be considered fundamental, and this course, I think, is wrong.

The denial of action at a distance in this sense does not carry with it the acceptance of the notion of an ether. The latter concept was effectively that of an elastic solid capable of transmitting transverse waves with a constant velocity, and has broken down under later work. But the ideas of position co-ordinates and time, and of the electric and magnetic forces associated with them, arise of themselves, quite independently of the assumption of a quasi-material substance filling space. Our knowledge of electromagnetic phenomena indicates that they are related by differential equations, which in turn imply and explain the properties of light. The question of an ether does not arise.

The principle of causality now becomes the aggregate of all scientific laws, whether already known or awaiting discovery. To accept it implies a hope that we may some day know all laws; but that day is still distant. As a working rule it may be valuable for its psychological effect, but there is so far no definite reason for believing it true, and science can get on quite well without it.

The words *cause*, *effect*, and *because* are on a different footing, and have nothing to do with a general principle of causality. If a scientific law involves a number of variables, then a knowledge of all but one of them determines that one. We say that it has a certain value *because* the others have certain values. The notions of cause and effect involve rather more than this; there is an asymmetry about them that is absent from the word *because*. Thus we may say either that a triangle has the angles at the base equal because it is isosceles, or that it is isosceles because the angles at the base are equal. When we speak of a cause and an effect, we pick out the one as the cause and the other as the effect, and they cannot be interchanged. The distinction seems to be one of time; the events under discussion are connected by a scientific law, and we pick out the earlier and call it the cause, and the later the effect. There is no distinction of cause and effect for contemporaneous events. The definition of simultaneity on the

principle of relativity makes it possible to generalize this. We have seen that two events that are simultaneous for one observer are not necessarily simultaneous for another; but if two events are specified by position co-ordinates and time, (x, y, z, t), (x', y', z', t') for any observer, and we consider the quantity

$$r^2 = c^2 (t - t')^2 - (x - x')^2 - (y - y')^2 - (z - z')^2,$$

then r^2 has the same value for all observers in the same universe. If it is positive, we can say that the one event is before or after the other, and it is possible for a message travelling from one place with a velocity less than that of light to reach the other place in time $| t - t' |$. If t' is the greater we say that the event (x', y', z', t') is the later of the two. If r^2 is negative, a message would have to travel with a velocity greater than that of light, and no such velocity is known in physics. Then we say that neither event is before or after the other, and in fact we can find velocities of the observer that would make them simultaneous. Thus events are arranged in what Robb calls a *conical* order in terms of the invariant relations of *before* and *after*, where we say now that one event is before or after another if it is before or after it to *all* observers. Then we can say that if there is a law connecting two events, the earlier in this sense is the cause and the other the effect, and this is a definition that applies to all systems of reference. If two events are connected, but neither is before or after the other, we may use the word *because*, but we cannot say that either is the cause and the other the effect. In such cases we can, of course, usually trace both to some cause earlier than either.

Related to the question of repetition is the case where the thing under observation is destroyed by the act of observation. Thus when we analyse a chemical compound the final products are not the same as the original compound; when we observe light it is absorbed in the eye and not re-emitted; and when a neurosis is psychoanalysed the patient recognizes

the relation of the symptom to his early conflicts, which are no longer of practical importance to him, and the neurosis disappears. There is again no difficulty in principle when we recognize that our data are sensations. The chemical compound, light, and the neurosis are all concepts designed to explain sensations, and there is no difficulty about supposing that the concepts cease to exist when the sensations they were designed to explain no longer exist. We do need, of course, to recognize the memory of previous sensations among our data.

10·3. Some reference may be made here to the practice in mathematical physics of "neglecting small quantities" and arguing by "orders of magnitude". Both methods are almost universally accepted by physicists; both are looked upon somewhat askance by pure mathematicians; and both are completely unintelligible to the man in the street, to whom the journalistic expression "mathematical accuracy" implies an entirely erroneous idea of what mathematics means. Thus the problem of "squaring the circle" still has its devotees; some of the uninitiated try to solve it by methods that are known to be incapable of solving it, and others repeat the legend that the problem is still unsolved. I remember once seeing a claim in a popular science journal that, though π has been evaluated to a large number of decimal places, all such estimates are wrong because they are not exact; the author proceeded to prove to his own satisfaction that it was exactly equal to 3·125. This is an extreme case; but there was apparently a publisher willing to pay for printing the article, and presumably there was a public willing to buy the journal. We need not take this proposition seriously; we need only notice that it has an emotional value expressible in terms of hard cash. The apparent precision of the number 3·125 was the attraction; the fact that π is known to be between 3·14159 and 3·14160 was considered irrelevant. On a somewhat higher level, we have the candidate in the Mathematical

Tripos who attempts to do a problem in small oscillations without neglecting the squares of small quantities till the very end. He never gets the right answer (even if he gets an answer at all), because he always makes a mistake in algebra. We have also the candidate who can do a complicated factorization but cannot prove the simplest inequality. In this we must detect an inherent tendency to trust the word *equal*, but a suspicion of *greater than* and *less than*, which is scarcely exceeded by that directed against *approximately equal*.

In the sense understood by the man in the street, exactness has almost disappeared from the subject-matter of modern pure mathematics. It survives in projective geometry, which is really the study of sets of algebraic equations, and in the identification of high prime numbers. Modern analysis deals with infinite series and the behaviour of integrals and differential coefficients, all of which involve the notion of a *limit*. Thus the sum of an infinite series whose nth term is u_n is defined as the limit, if any exists, of the sum of the first n terms when n becomes indefinitely large. The criterion that the sum may be S is that, if we choose any positive quantity ϵ, however small, we can find a value of n such that for *all* values of m greater than n, the sum of the first m terms differs from S by less than ϵ. While the sum S appears as an exact value, it is the result of a limiting process, which depends essentially on a recognition of the meanings of *greater than* and *less than*. The solutions of most of the differential equations of physics are expressible as series possessing sums so defined, and their numerical values can be found by actually computing the series, term by term, till the desired accuracy is obtained.

The physicist often looks at the summation of series from a different standpoint. He is not interested in the fact that the sum of n terms of the series has a definite limit when n becomes indefinitely large; he has not the slightest intention of computing more than a certain finite number of terms. The existence of the function associated with the series is usually known already, and what the physicist needs is to evaluate

it with the requisite degree of accuracy without a prohibitive amount of labour. The mathematical property of convergence is neither a necessary nor a sufficient condition for physical utility. Thus if we consider the exponential series

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots$$

and put $x = -1000$, we have a series that converges in the mathematical sense. But the terms increase numerically up to $n = 1000$, and the actual computation would be hopeless. Actually, of course, one writes

$$e^{-1000} = 1/10^{1000 \log_{10} e}$$

and computes $\log_{10} e$ from some such formula as

$$1/\log_{10} e = \log_e 10 = 3 (\log_e \tfrac{3}{2} + \log_e \tfrac{4}{3}) + \log_e \tfrac{5}{4}.$$

Convergence is therefore not a sufficient condition for utility. Nor is it necessary, for if we consider the series

$$e^{x^2} (1 - \operatorname{erf} x) = \pi^{-\frac{1}{2}} \left(x^{-1} - \frac{1}{2} x^{-3} + \frac{1 \cdot 3}{2 \cdot 2} x^{-5} - \frac{1 \cdot 3 \cdot 5}{2 \cdot 2 \cdot 2} x^{-7} + \ldots \right),$$

the series on the right is always divergent. But it can be shown that the sum of the first n terms always differs from the function on the left by less than the last term retained. If x is large, the terms decrease to a minimum, and the smallest may be within the range of accuracy required. Thus for $x = 3$, the fourth term is about $\frac{1}{400}$ of the first. Such series are called *asymptotic*, and have been extensively studied in modern pure mathematics. But whereas the tendency of the pure mathematician is to consider convergence as the generally important property, and the asymptotic property as a makeshift, physical utility makes the asymptotic property valuable and convergence unimportant. But the real test for physical utility is that the sum of the first n terms (where n is preassigned, usually does not exceed 10, and often is 1), shall differ from the function represented by less than the limits of error permitted by other considerations. If this condition is

satisfied it is no concern of the physicist's how the later terms of the series behave. If it is not satisfied he will have recourse to numerical solution of the differential equation.

The "neglect of small terms" in a differential equation implies that the solution is in error to some extent, which will depend on the actual magnitude of these terms. What is certain is that the solution will remain approximate throughout a certain range of the independent variable; the smaller these terms are, the longer the range will have to be before their integrals become large enough to affect the accuracy of the approximation to a given extent. In some cases we can prove rigorously that they will never do so: in others we cannot. It seems to me that the general theory of the degree of accuracy of these approximations is an important and almost unworked field of pure mathematics. The physicist's method is to solve the problem first by neglecting them, and to substitute the result in the small terms to verify that they do remain small.

The use of "orders of magnitude" is a further departure from popular standards of accuracy. It usually consists essentially in the principle that if x varies from a to b, a function $f(x)$ varies from $f(a)$ to $f(b)$, and we can replace its derivative $f'(x)$ by its mean value $\frac{f(b)-f(a)}{b-a}$. If $f'(x)$ is continuous this is true for some value of x between a and b; but the method goes further. If we have a differential equation we may carry out operations of this type on both sides of it and reduce the equation to a single algebraic equation. The result is necessarily inaccurate. Its utility is essentially in carrying out preliminary tests on a theory. If we get an agreement within a numerical factor of 5 or so we may say that the theory is worth closer examination; if the two sides of the equation so obtained differ by a factor of 1000 or more, we consider that further investigation is unnecessary. There may be physical grounds, in a particular problem, for contrasting two hypotheses directly, and then the method of orders of

magnitude will enable us to reject one and retain the other without the trouble of carrying out an accurate investigation.

These methods hardly arise, I think, in the *establishment* of a physical law. They are concerned with the investigation of the competence of different causes to produce a given effect, the laws being already known.

The term "order of magnitude", in the physical sense, means rather more than it does in modern pure mathematics. Thus the pure mathematician may write an equation

$$f(x) = \phi(x) + O(x^2),$$

where $f(x)$ and $\phi(x)$ are two known functions of x, and he will say that their difference is of the order of magnitude of x^2. He means that when x tends to zero, the ratio $\{f(x) - \phi(x)\}/x^2$ tends to a finite limit or zero. The limit may be 1000. A physicist in such a case would not say that $f(x) - \phi(x)$ is of the same order of magnitude as x^2, for he probably wants the actual values of the functions when x is different from zero, and if the limit is a large number the utility of the approximation may be vitiated. The physicist's meaning is more restricted in one way, though less precisely defined in another. He may say that two quantities are of the same order of magnitude when there is no question of a limit; thus the masses of Jupiter and Saturn are of the same order of magnitude. Two quantities may be said to be of the same order when their ratio does not exceed 10; and the justification of the method is that the ratio of the mean values actually compared in the reduced equation is really a numerical constant arising in the solution, and that in practice such constants hardly ever do exceed 10. Exceptions sometimes arise: thus the condition that turbulence may persist in the flow of a fluid in a pipe involves a numerical constant of the order of 1000, but that is really because the solution of the problem involves not one equation, but a family of four differential equations, three of the second order and one of the first. Thus there may be such numerical constants as $7!$ or 5040.

CHAPTER XI

OTHER THEORIES OF SCIENTIFIC KNOWLEDGE

I have seen all the works that are done under the sun; and, behold, all is vanity and vexation of spirit. Eccles. I. 14

A preliminary explanation is needed before entering on the topics of this chapter. The theories considered here are selected on account of their relation to the general aim of this book, which is to systematize the processes actually employed in the acquisition of knowledge by experience. They have in common, in my opinion, the feature that if they were accepted as practical rules of working they would make this acquisition impossible. In some cases they were expressed by their authors some time ago, and I am not in all cases in a position to know whether the respective authors still hold the views in question. For my purpose, however, it is the theories themselves that matter, rather than the personal question of whether the individual authors still hold them; for in fact each theory still certainly has a number of professed adherents.

11·1. *The statistical theory of probability.* In the present work probability is regarded throughout as a property of the relations between propositions. Like such notions as force, interval of time, distance, electric current, colour, pitch of sound, and so on, it is immediately recognizable by consciousness in suitable circumstances. Like them also its treatment can be made quantitative, and its specification can thereby be made enormously more precise. The original meaning, however, is never lost. If we ignore it we are deliberately neglecting a piece of knowledge that we have, and are therefore restricting the application of scientific method. This is more serious in the case of probability than with the other

concepts mentioned, because it is not the subject-matter of a branch of science; science is a branch of the subject-matter of probability. To ignore probability is to ignore the problem of scientific inference and to deprive science of its chief reason for existence.

Many writers, following the late John Venn*, have attempted to avoid the notion of probability as a primitive concept by trying to define it in terms of the composition of samples. Venn considered that the notion of probability presupposes a series, the terms of which are indefinitely numerous and represent the cases of an attribute ϕ. From these one can pick out a smaller class, the members of which possess the further attribute ψ. If then we have chosen m members in all, and l of them belong to the smaller class, the probability of ψ given ϕ is defined as the limit of l/m when m becomes indefinitely great. The form of this definition restricts the field of probability very considerably. As a matter of simple fact, when we speak of probability we do not consider an indefinitely large number of trials. In many cases, such as when we speak of the probability that the solar system was formed by the disruptive approach of two suns, or that the stellar universe is symmetrical, the idea of even one repetition is out of the question. Yet these are precisely the cases where the notion of probability is most valuable.

But actually Venn's definition suffers from a drawback that deprives it of all application whatever. If a definition is to be of any use it must imply a test, and we must be able to carry out that test. On the *a priori* view, when we say that the probability that a penny will come down heads is $\frac{1}{2}$, we make an immediate judgment. On Venn's view we must throw it an infinite number of times and take the limit of the ratio of the number of heads to that of all throws, and nobody has had, or ever will have, time to do it. There is no case where the value of the probability, on Venn's definition, is known, or even where it is known to exist.

* *Logic of Chance*, pp. 162 et seqq.

We must remember that this view is designed to avoid the need to treat probability as an undefined concept with *a priori* laws of its own. The undefined concept view gives a justification for the opinion that a large sample will probably be approximately a fair one; if we reject this view we also reject the justification that it gives, and must be prepared to find a new one. The question at issue is whether, apart from the *a priori* view, there is any reason to believe that the ratio considered in Venn's view tends to any limit whatever. To say that it does is essentially an assertion about the result of an experiment that nobody has ever tried, or ever will try.

It can be seen easily that, with any value of the probability whatever, other than 0 or 1, it is possible to have selections that do not give any limit for the ratio. For if the ratio is to tend, when m is large, to any limit between 0 and 1, the numbers of things possessing and not possessing the attribute ψ are both infinite. We cannot take the actual ratio of the whole number of ψ's to the whole number of ϕ's to express the probability, for both numbers are in fact infinite and their ratio is indeterminate*. The method of proceeding to the limit is essential to the definition. But if at any stage we are able to select either a ψ or a not-ψ, it is possible to make the limit anything whatever, or there may be no limit at all. If, for instance, whenever a ψ occurs we write 1, and whenever a not-ψ occurs we write 0, l/m will be the mean of the first m terms in the series obtained. If they occur in such an order as to give the series

1011000011111111.....

where the number of digits in any block after the first is equal to the number of digits that have occurred previously, the ratio is about $\frac{2}{3}$ at the end of each block of 1's, and about $\frac{1}{3}$ at the end of each block of 0's. It therefore tends to no

* R. A. Fisher, with what looks like the courage of despair, says that in a "hypothetical infinite population" the ratio is perfectly definite. Cf. *Phil. Trans.* 222 A, 1922, p. 312.

limit whatever. Again, in one selection we may get the series

1010101O.....

which gives the limit $\frac{1}{2}$; but from the same class we could make the selection

1001001OO.....

which gives the limit $\frac{1}{3}$. The numbers of available 0's and 1's being both by hypothesis infinite, there is no possibility of exhausting either, so that such series are in fact possible. It is therefore possible for Venn's ratio to tend to the wrong limit, or to give no limit whatever. The very existence of the probability on Venn's definition requires an *a priori* restriction on the possible orders of occurrence of ψ's and not-ψ's. No supporter of this view has succeeded in stating the nature of this restriction, and even if it were done it would constitute an *a priori* postulate, so that this view involves no reduction of the number of postulates involved in the treatment of probability as an undefined concept with laws of its own.

The difficulties become worse when we attempt to combine probabilities, for then we have to face an indefinite repetition of infinite series. This is called by Venn the use of *cross-series*, and forms an important part of his theory of inference. It is necessary, for instance, in giving a meaning to the proposition connecting the probabilities of a proposition referred to different data,

$$P(p \,.\, q \mid h) = P(p \mid q \,.\, h)\, P(q \mid h).$$

For an infinite series is necessary to give an account of $P(p \mid q \,.\, h)$, which is the limit derived from the frequency of the truth of p among entities satisfying q and h. Such entities are, however, only a part of those that satisfy h. Thus to establish a meaning for the numbers $P(p \,.\, q \mid h)$ and $P(q \mid h)$ we must consider all entities satisfying h, whether they satisfy q or not. Thus further series must be constructed to show how often q is actually true, and this requires, according to Venn,

an infinite number of series of entities all satisfying h, so that we can examine in one direction to find the frequency of p given q and h, and in the other direction to find those of q given h and of $p \,.\, q$ given h. Thus the difficulty of obtaining enough terms, an acute practical point in the simple case, is here intensified. Further, there is no more reason to believe in the existence of limits in this case than there was in the other; and the opinion that the limits, if they exist, will satisfy the relation is justified only if the samples are made according to some special rule. The difficulties are merely complicated and not removed by the use of cross-series; and the statistical theory of probability becomes a network of begged questions.

There is a question of the theory of probability, treated as an undefined concept, that is related to the question of the existence of Venn's limit. If the probability of ψ given ϕ is r, and is the same however many instances have been examined, what is the probability that when the sample becomes indefinitely large the ratio l/m does tend to the limit r? To say that it does so means that for any quantity ϵ, however small, we can find a number m_0 such that, for *all* values of m greater than m_0, l/m is between $r \pm \epsilon$. What is the probability of this proposition? It has not, so far as I am aware, been evaluated, and a determination would be interesting. It is not rigorously unity, since it has already been shown that there are possible samples that do not satisfy the proposition. It may, however, differ from unity either finitely or infinitesimally. If the difference is finite, the Venn definition loses the last of its justification from the undefined concept view. If it is infinitesimal, we might, if we thought the definition worth saving, save it at the cost of admitting infinitesimal probabilities different from zero.

11·2. *Keynes's theory of probability.* To those already familiar with J. M. Keynes's *Treatise on Probability* (1921) it will be obvious that the point of view of the present work is very

similar in some places and very different in others. The task of comparing the developments explicitly, point by point, would be too formidable, but could for the most part be achieved by the reader sufficiently interested to carry out a direct comparison. Keynes agrees with me in regarding probability as an undefined concept, really following De Morgan and Jevons, with a series of earlier writers going back to Leibnitz. He differs from the earlier writers, and from me, in refusing to admit that all probabilities are expressible by numbers. This amounts to denying the postulate of the present theory, that of any two probabilities one is greater than, equal to, or less than the other; or the equivalent, that of any three unequal probabilities one is between the other two. Granting this proposition, it has been proved in this work that probabilities can be uniquely associated with numbers. Keynes's alternative is something like the view that probabilities resemble places on the earth's surface; we might say that New York and London are both between the North and South Poles, but neither New York nor London is between the other and the North Pole. It seems to me that all probabilities actually are comparable and that Keynes is merely creating difficulties. He manages to preserve the form of the probability of the disjunction of two propositions by defining addition in terms of it; that is, the proposition

$$P(p \,.\, q \mid h) + P(p \,.\, \sim q \mid h) = P(p \mid h),$$

which to me is a law connecting numerical estimates of probability, is to Keynes the *definition* of addition, and the terms in it may not be numbers at all. Similarly the law 2·32 (4) is converted into a definition of multiplication. The mathematical development remains much the same; the only question is whether the results mean anything. Thus on Keynes's views probabilities might be complex numbers; and then it is possible that inequalities involving products, which are true for real numbers, may break down, and arguments based

on the approach of probability to certainty with repeated verification may fail. But my real objection to Keynes's postulate is that it is one of those attempts at generality that in practice lead only to vagueness.

It might be held that, since different people do appear to assess probabilities differently, Keynes's postulate might fit the assigned probability instead of the true probability. But I do not think that this is the case. We know people who appear to assess all probabilities at either 0 or 1; we know others who seem to assess them all at $\frac{1}{2}$, whatever the available evidence; and there may be some who assign the probability 1 to their own hypotheses and $\frac{1}{2}$ to all those of other people, unless of course the latter happen to contradict their own, when their probability is 0. But such estimates do not follow the quantitative rules connecting the probabilities of propositions referred to different data, and can only be understood by introducing psychological considerations. I think the correct attitude to them is that they are simply wrong, just as it is possible to get a wrong answer in solving an algebraic equation.

For some other comments on Keynes's work I refer to *Nature*, Feb. 2, 1922, 132–133.

11·21. There is just a possibility that probabilities may in certain circumstances require for their expression more numbers than the real numbers. Just as the real numbers are more numerous than the rational fractions, it is possible to define continuous series with more members than the real numbers, and yet satisfying the condition that of any three members of the series one is between the other two. But this is a degree of generality that has not yet required recognition.

11·3. *Phenomenalism.* This theory of knowledge may be defined by the rule that nothing is to be supposed to exist that cannot be reduced to descriptions of sensations. It may be traced back to the mediaeval writer William of Ockham, who said, "Entities are not to be multiplied beyond neces-

sity", and as such was probably a reaction against the disposition of primitive man to postulate an independent god or demon as a cause for everything he could not understand. In its modern form it is effectively due to Ernst Mach and Karl Pearson, whose discussions of the bases of mechanics led more than anything else to the recognition of the need to define force and mass in terms of actual experience, so far as possible, and to the dropping of such ideas as absolute position and ether. Having myself started from the phenomenalist position, I must express my great indebtedness to these writers, but I consider that the pure phenomenalist attitude is not adequate for scientific needs. It requires development, and in some cases modification, before it can deal with the problems of inference. We must, as has been said already, always distinguish between sensations actually experienced and those inferred from other sensations. The former can be described; the latter can only be inferred with greater or less degrees of probability. Mach hardly considers the question of probability; Pearson does not go beyond Laplace's theory. It has been shown here that a requisite of any satisfactory theory of inference, as actually carried out in scientific work, is a recognition of the high prior probability of the simple law. There is no harm in concepts that cannot be defined as classes of sensations, provided that a few of them will help in describing a large number of sensations. This is the test of the scientific validity of a concept; philosophical reality has nothing to do with it. An electron, for instance, is a valid scientific concept; I think that it is merely playing with words to say that it is a class of sensations, or that it can be described in terms of sensations. The same applies to the matter at the centre of the earth, or to the state of the earth just after its formation; both enable us to co-ordinate sensations actually experienced and are therefore admissible concepts.

11·4. *The theories of Russell and Whitehead.* Mr Bertrand Russell, in *Mysticism and Logic* (1917), tries to tackle the

problem of actually defining objects in terms, not exactly of sensations, but of *sense-data*, which are effectively sensations with the errors of observation removed. Physical objects still cannot be adequately defined as the class of those sense-data that, in ordinary language, would be said to be perceptions of it, for then the object would change when new aspects of it are observed, and this is not to be allowed. Therefore he considers the object defined in terms of all possible aspects of it; these aspects are called *sensibilia*, and resemble sense-data in everything except that the majority of them are not perceived. A physical object is then a class of sensibilia.

From the practical scientific standpoint the weakness of this attitude is that we do not know what the sensibilia are like. An object, on this theory, could never be described until we had a knowledge, by experience, of all its aspects, perceived and unperceived, and this is inherently contradictory. Even the perceived sensibilia, or sense-data, cannot be described in terms of sensations until we have some rule for removing the errors of observation. The unperceived ones are necessarily never known directly, but have to be inferred from the perceived ones; and this can be done only by using the laws of physics, inferring the nature of the object, and then proceeding to the unperceived sensibilia. The physical object and the laws of physics are anterior in knowledge to the sensibilia, and Mr Russell's theory, whether it is logically consistent or not, is not a theory of scientific knowledge.

In Prof. Whitehead's theory* events, instead of sensibilia, are the fundamental entities. Each event contains other events, so that we can specify series of events such that each event in a series surrounds all after it. The limit of such a series is a point-event, and it is to such point-events that the laws of physics are supposed to apply. But the notion of a limit requires an infinite class, and an infinite class of observations is impossible in practice.

* *An Enquiry into the Principles of Natural Knowledge*, 1919.

We may say that it is never possible to construct a valid theory of knowledge that involves the use of infinite classes of empirical data. The objection is similar to that given by Poincaré* in his criticisms of Cantor's theory of infinite numbers. Poincaré argued that it is impossible to assert anything about a class, and in particular anything about the number of its members, until every member of the class has been defined in words; and as only a finite number of entities can ever be defined in words, it is impossible to know anything about an infinite class, so that there can be no knowledge of infinite numbers. The argument, as it stands, is not valid against Cantor's theory, for in order to make an assertion about a class it may not be necessary to have definitions of all the members separately; often a general proposition about all members can be asserted or postulated, and is enough for the purpose. Poincaré, indeed, seems to have overlooked the fact that if his argument were sound it would also destroy the whole theory of infinite series and of differentiation and integration; thus little would be left of higher pure mathematics. Thus the convergence of a series depends on the proposition that the sums of the first n, $n + 1$, $n + 2$, ... terms, for some value of n, all differ from a certain number, called the sum of the series, by less than a fixed quantity ϵ. These sums are infinite in number, and hence it would be impossible, if Poincaré's assumption were granted, ever to prove that a series is convergent. This result is, of course, quite unacceptable. But the argument would go even further than this. Nobody has had time in his life to construct definitions of every member of a class of a million members; and as a number is merely a property of a class, it should be impossible to prove that, for instance,

$$1\,000\,001^2 = 1\,000\,002\,000\,001.$$

Thus the argument would also invalidate most of arithmetic. If therefore we believe that the propositions of arithmetic

* *Science et Méthode*, 1908, 192–214.

have some meaning and are true, we cannot accept Poincaré's objection to the theory of infinite numbers.

But while the argument is wrong in this case, it is clearly valid when our only source of information about the members of a class is empirical; for the total number of observations a person can make in his life is finite, and hence his experience alone can never tell him anything about all the members of an infinite class of entities. Any proposition about such a class, or about all its members, is necessarily either wholly *a priori* or else an inductive generalization, and neither known directly nor obtainable from experience by the principles of pure logic alone. The fundamental data of any branch of science must consist of a finite number of observational results and some *a priori* postulates.

One consequence of this is that we can never prove the existence of a limit to which a series of entities known by experience may tend, for in order to establish the existence of such a limit we should need to have knowledge that an infinite number of such entities are within a definite distance of that limit. This by itself would not be a fatal objection to any such theory, for there seems to be no possibility of constructing a theory of knowledge without some assumptions, and it may be considered that in the case in question certain conditions are satisfied under which the existence of a limit is known *a priori*. But what is fatal is that in physical problems we do not merely want to know that the limit exists; we want its value according to some definite system of measurement, and that value can never be known *a priori*; indeed, if it were, there would be no need to make measurements at all. Thus if a limit is ever used in a scientific theory, its value and all propositions about it are neither *a priori* nor known by experience, and therefore are not primitive propositions that can be used in a theory of knowledge based on experience. It is seen that this consideration rules out at once the statistical definition of probability, with the theories of Russell and Whitehead just mentioned.

APPENDIX I

PROBABILITY IN LOGIC AND PURE MATHEMATICS

By convention it has been decided that if the proposition p implies q, the probability number $P(q \mid p) = 1$. The word *implies* is used in the ordinary sense, namely, that if p is true, then q is true. This is the definition given by Whitehead and Russell. There is, however, a slight difficulty. Whitehead and Russell prove the propositions

$$\left.\begin{array}{r} p \text{ implies that } q \text{ implies } p, \\ \sim p \text{ implies that } p \text{ implies } q. \end{array}\right\} \qquad (1)$$

These are often read "a true proposition is implied by every proposition" and "a false proposition implies every proposition, true or false". But when we analyse these propositions in terms of the definition, they become

$$\left.\begin{array}{l} \text{If } p \text{ is true, then if } q \text{ is (also) true, } p \text{ is true.} \\ \text{If } p \text{ is untrue, then if } p \text{ is true, } q \text{ is true.} \end{array}\right\} \qquad (2)$$

The first is now seen to mean simply that additional (true) information does not contradict a proposition already known to be true; its paradoxical appearance is gone. It is expressed in our rule

$$P(p \mid p \cdot q) = 1. \qquad (3)$$

The second, on the other hand, does not enable us to infer q without the knowledge that p is both untrue and true; and this circumstance fortunately never arises. But formally it appears to require

$$P(q \mid p \cdot \sim p) = 1; \quad P(\sim q \mid p \cdot \sim p) = 1, \qquad (4)$$

and therefore

$$P(q \vee \sim q \mid p \cdot \sim p) = 2, \qquad (5)$$

whereas no probability can exceed 1. Similarly we could write $\sim p$ for q in (3) and get

$$P(p \mid p \cdot \sim p) = 1, \tag{6}$$

and replacing p by $\sim p$

$$P(\sim p \mid p \cdot \sim p) = 1. \tag{7}$$

It seems that contradictions are inevitable if we adhere to these propositions and allow contradictory propositions to appear among the data simultaneously. I think the correct convention in these circumstances would be that the probabilities are simply indeterminate. Thus we have

$$P(p \cdot \sim p \cdot q \mid h) = 0$$
$$= P(q \mid p \cdot \sim p \cdot h)\, P(p \cdot \sim p \mid h),$$

and $P(p \cdot \sim p \mid h) = 0$. Hence $P(q \mid p \cdot \sim p \cdot h)$ is of the form 0/0 and therefore is indeterminate.

So far as the theory of scientific method is concerned, the point is, of course, purely academic. Our estimates of probability are always to be based ultimately on *a priori* principles and sensations, which are never mutually contradictory, so that the difficulty can never give any trouble in practice.

It might be suggested that the statement "p implies q" means more than "if p, then q", and requires an actual proof that the relation of implication holds for the propositions p and q; until this is given, the probability of q given p is less than unity. I think that this is a wrong attitude. Consider some undemonstrated proposition of pure mathematics, such as Fermat's last theorem, or the proposition that the thousandth decimal in the expression for e is zero. The data in each case are perfectly definite. In the latter case it is known how the proposition could be tested if anybody was sufficiently interested to do the work; in the former all the powers of the natural numbers are perfectly definite, and it is only a question of whether actually, with a value of r greater than 2, two numbers x and y exist such that $(x^r + y^r)^{1/r}$ is a whole number. Each proposition is simply true or false on the data

of pure mathematics themselves; a proof does not affect their truth-values, but merely finds out what they are. The probability of Fermat's last theorem, given the data of pure mathematics, is therefore either 0 or 1; we simply do not know which. The proposition "it can be proved that Fermat's last theorem is true", on the other hand, is different from the proposition "Fermat's last theorem is true", for it introduces the question of the possibility of proof, which is a question of the capabilities of the human mind, and a legitimate field for scientific investigation based on experience. In view of the efforts that have been made to prove the theorem, we may say that the probability of this proposition is small, though not absolutely zero.

APPENDIX II

INFINITE NUMBERS

The following remarks are not intended as a full account of the modern theory of infinite numbers. This book is meant mainly for theoretical and experimental physicists, and for their purposes a brief summary is probably all that is needed. If more is required, G. Cantor's *Transfinite Numbers*, or Littlewood's *Elements of the Theory of Real Functions*, may be read; a full account is in Whitehead and Russell's *Principia Mathematica*. A glance inside is worth while, as the inside is even more impressive than the outside.

The fundamental notion involved in number is that of comparison of classes. If we have two classes α and β, such that they can be arranged so that to every member of α corresponds one member of β, different α's corresponding to different β's, then the number of members of α is less than or equal to that of β, and that of β is greater than or equal to that of α. If the classes can also be arranged so that to every member of β corresponds one of α, then the classes are said to be equal in number, and an arrangement can be found so that each member of either class corresponds to one of the other, none at all being left over. The smallest infinite number is the number of the positive integers; this is called $\aleph_0$. We can prove that $\aleph_0$ is also the number of the rational fractions. For we can arrange the rational fractions thus:

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \frac{1}{5}, \frac{5}{1}, \frac{1}{6}, \frac{2}{5}, \frac{3}{4}, \frac{4}{3}, \frac{5}{2}, \frac{6}{1}, \frac{1}{7}, \frac{3}{5}, \ldots\ldots .$$

Here we first of all group together those fractions with the sum of the numerator and denominator the same, and arrange the groups so that this sum is greater in the fractions of one group than in those of any earlier group. In each group we place the fractions in order of increasing numerator. This arrangement includes every rational fraction, and they are

put in a definite order, so that every fraction is reached in a finite number of steps from the beginning. The positive integers 1, 2, 3, ... can therefore be placed against them. A one-one correspondence is therefore established between the rational fractions and the positive integers, and the two classes therefore have the same number.

It can be shown similarly that the number of numbers that are the roots of algebraic equations with rational coefficients is $\aleph_0$. For an equation may first be multiplied by the lowest number that will clear it of fractions. For each equation we can take the sum of the absolute values of the coefficients and the degree, and we can arrange the equations in groups according to increasing values of this sum. The number of equations in each group is finite, and the total number of their roots is also finite. Thus we can arrange both the equations and their roots so that every member is reached in a finite number of steps from the beginning; the number of algebraic equations, and the number of their roots, are therefore both $\aleph_0$.

Similarly the number of differential equations of finite order and degree, such that each coefficient is capable of $\aleph_0$ values, is $\aleph_0$. For we can begin by arranging the $\aleph_0$ values of each coefficient so that they correspond to the whole numbers, and then replacing the values by the whole numbers themselves. This gives a new class of differential equations with the same number as the first. Now rationalize each equation. Then form for each equation the sum of the order, the degree, and all the coefficients. Arrange the equations in groups, according to increasing values of this sum. The number of equations in each group is finite, and therefore we can arrange the equations so that every member is reached in a finite number of steps from the beginning; and the total number of these equations is infinite. Hence the number is $\aleph_0$.

If we have two classes of numbers m and n, we can form pairs of things one from each class. The possible ways of forming such pairs are mn in number. This is taken as the

definition of the product of two numbers. We can prove that the product of $\aleph_0$ by any finite number or by itself is also $\aleph_0$. Considering the latter proposition, $\aleph_0 . \aleph_0$ would mean the number of pairs of the form (x, y), where x and y are whole numbers. We can arrange these pairs in groups for which $x + y$ has the same value, and then arrange the pairs in each group in order of x increasing. In this way the pairs are arranged in order so that each is reached in a finite number of steps from the beginning, and their number is infinite. Hence their number is $\aleph_0$. Thus $\aleph_0 \cdot \aleph_0 = \aleph_0$.

Suppose next that we have two classes of numbers m and n. With any member of the first class we can associate one of the second class in n ways; with another member of the first class we can associate one of the second in n ways (repetitions being allowed). Then we can say that the number of ways of covering the two together by members of the second class is n^2. If we consider every member of the first class associated with every member of the second, the whole number of ways of carrying out such pairings is called n^m. This operation is called exponentiation. In particular if m numbers are to be assigned and each of them can take n values, the total number of ways of assigning values to all of them is n^m.

The number of decimals to the base n is $n^{\aleph_0}$. For in each decimal there are $\aleph_0$ places to be filled, and n possible numbers 0, 1, 2, ... $n - 1$ can be placed in any place, irrespective of what numbers are in any other place. If we assume that any real fraction, rational or irrational, can be expressed as a decimal with any base, it follows that the number of real fractions is $n^{\aleph_0}$, where n may be any whole number greater than 1. Also, since our fraction can be reduced in particular to the base 2, we have

$$n^{\aleph_0} = 2^{\aleph_0} = C,$$

say. This number C, the number of real fractions, is called also the number of the continuum.

It can be proved that if n is the number of a class, 2^n is

always greater than n. We need this result especially for the case where $n = \aleph_0$. Suppose, if possible, that $2^{\aleph_0}$ was equal to $\aleph_0$. This would mean that all real numbers could be so arranged that they corresponded one by one to the whole numbers in ascending order. Then imagine each converted to a decimal to the base 2. In each place the digit is either 0 or 1. We can now construct a decimal that differs from the first of the series in the first place, from the second in the second place, and so on. This decimal will then differ from every member of the series that was supposed to include all. It follows that $2^{\aleph_0} > \aleph_0$.

It follows at once that $2^{\aleph_0} > \aleph_0{}^n$, where n is any finite whole number. For by repeated application of the result that $\aleph_0{}^2 = \aleph_0$ we can show that $\aleph_0{}^n = \aleph_0$.

The same set of things may be arranged in different ways. If an arrangement is such that each member of the series has an immediate successor, we say that the series is *well-ordered*. Thus the whole numbers 1, 2, 3, ... in ascending order of magnitude constitute a well-ordered series; for to each number there corresponds a "next" such that the latter follows it and there is none between. The rational fractions, or the algebraic numbers, in ascending order, are not well-ordered, because between any two there lie an infinite number of others. These series are called *dense*. The whole numbers in ascending order are said to be of ordinal type ω; the rational fractions or the algebraic numbers between 0 and 1 in ascending order, omitting 0 and 1 themselves, are said to constitute a series of ordinal type η. It is doubtful whether every class can be arranged so as to be well-ordered; the doubt extends to the continuum itself. The continuum of real numbers between 0 and 1, omitting the extremes, is said to be of type θ.

Little progress has been made with the theory of the ratios of infinite numbers. It appears to be very doubtful whether such ratios could be defined so as to satisfy the ordinary rules of fractions; we should wish, that is, to be able to multiply or divide numerator and denominator by the same

number without altering the fraction. But if we try to do so here we get, for instance,

$$\frac{1}{\aleph_0} = \frac{\aleph_0}{\aleph_0^2} = \frac{\aleph_0}{\aleph_0} = 1.$$

In the circumstances it seems undesirable to admit them to the theory of scientific method, at least until they have some recognized status in pure mathematics.

The number of functions of a real variable is C^C. For to every value of the variable may correspond any of C values of the function; and if the variable can assume any value within a continuum, it has C possible values. Hence by definition the number of functions is C^C.

The number of continuous functions is C. For if a continuous function is assigned for all rational values of x it is determined for all other values. For each rational value of x, the function can take C values. Hence it can be assigned for all rational values in $C^{\aleph_0}$ ways. But

$$C^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0^2} = 2^{\aleph_0} = C.$$

The same is true for analytic functions, even when each coefficient can take only $\aleph_0$ values. For the number of such functions, n say, is less than or equal to C, since they form only a part of all analytic functions. But their number is $\aleph_0^{\aleph_0} \geqslant 2^{\aleph_0} = C$. Thus

$$C > n \geqslant C,$$

and therefore $n = C$.

APPENDIX III

THE ANALYTIC TREATMENT OF THE SINE AND COSINE

We define $\cos x$ and $\sin x$ by their expansions in series

$$y_0 = \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots, \tag{1}$$

$$y_1 = \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots. \tag{2}$$

Both series are absolutely convergent and differentiable term by term for all values of x. We see at once that

$$\frac{d}{dx}\cos x = -\sin x; \quad \frac{d}{dx}\sin x = \cos x, \tag{3}$$

and that both $\cos x$ and $\sin x$ satisfy the differential equation

$$\frac{d^2y}{dx^2} + y = 0. \tag{4}$$

If we multiply this by $2\frac{dy}{dx}$, the left side becomes a perfect differential and we infer that

$$\left(\frac{dy}{dx}\right)^2 + y^2 = \text{constant}. \tag{5}$$

Whether y is taken to be $\cos x$ or $\sin x$, we can substitute for its derivative from (3); hence

$$\cos^2 x + \sin^2 x = \text{constant}; \tag{6}$$

and putting $x = 0$ we see that the constant must be 1. Thus

$$\cos^2 x + \sin^2 x = 1. \tag{7}$$

When $x = 0$, $y_0 = 1$, $\frac{dy_0}{dx} = 0$, $\frac{d^2y_0}{dx^2} = -1$. Hence for a range of positive values y_0 is positive and decreases as x increases.

It must therefore either (1) vanish for a finite value of x, (2) begin to increase again without vanishing, or (3) tend to a finite limit less than 1 as x increases indefinitely. Alternative (2) implies that dy_0/dx vanishes for two values of x, 0 and another, and therefore d^2y_0/dx^2 must vanish between them. But this cannot be the case, because $d^2y_0/dx^2 = -y_0$ and y_0 is by hypothesis positive throughout the range. Alternative (3) implies that dy_0/dx tends to zero and therefore again d^2y_0/dx^2 must vanish for a finite value of x, which is again contradicted by the supposition that y_0 is always positive. Hence there is a value of x that makes y_0 zero. We call this $\frac{1}{2}\pi$. Evidently when $x = \frac{1}{2}\pi$, $\sin x = 1$ by (7). We have from (5) and (7)

$$\pm \frac{dy_0}{\sqrt{(1 - y_0^2)}} = dx, \tag{8}$$

whence by integrating and introducing the limits

$$\tfrac{1}{2}\pi = \int_0^1 \frac{dy}{\sqrt{(1 - y^2)}}. \tag{9}$$

Now consider the function

$$f(x_1, x_2) = \cos x_1 \cos x_2 - \sin x_1 \sin x_2, \tag{10}$$

and put $x_2 = x - x_1$. Then

$$f(x_1, x - x_1) = \cos x_1 \cos (x - x_1) - \sin x_1 \sin (x - x_1). \tag{11}$$

Differentiating with regard to x_1, we have

$$\begin{aligned} \frac{\partial}{\partial x_1} f(x_1, x - x_1) &= -\sin x_1 \cos (x - x_1) + \cos x_1 \sin (x - x_1) \\ &\quad - \cos x_1 \sin (x - x_1) + \sin x_1 \cos (x - x_1) \\ &= 0. \end{aligned} \tag{12}$$

Thus $f(x_1, x - x_1)$ is a function of x only. We can evaluate it by putting $x_1 = 0$, when we see that its value is $\cos x$. Now restoring x_2 we have

$$\cos (x_1 + x_2) = \cos x_1 \cos x_2 - \sin x_1 \sin x_2. \tag{13}$$

By differentiation with regard to x_1

$$\sin (x_1 + x_2) = \sin x_1 \cos x_2 + \cos x_1 \sin x_2. \tag{14}$$

Now replace x_1 by x and x_2 by $\frac{1}{2}\pi$. Then

$$\cos\left(\tfrac{1}{2}\pi + x\right) = -\sin x; \quad \sin\left(\tfrac{1}{2}\pi + x\right) = \cos x. \quad (15)$$

Replace x now by $\frac{1}{2}\pi + x$. Then

$$\cos(\pi + x) = -\sin\left(\tfrac{1}{2}\pi + x\right) = -\cos x;$$
$$\sin(\pi + x) = -\sin x. \quad (16)$$

Now replace x by $\pi + x$, and we have

$$\cos(2\pi + x) = \cos x; \quad \sin(2\pi + x) = \sin x. \quad (17)$$

Therefore the functions $\cos x$ and $\sin x$ have period 2π.

We have thus obtained from the analytic definitions the differential equation (4) satisfied by the cosine and sine; the relation (7) showing as a corollary that these functions cannot have absolute values exceeding 1 for real values of the argument; the addition formulae (13) and (14); and the periodic property (17).

LEMMAS

I. *Approximation to* $f(l) = {}^rC_l\, x^l\, y^{r-l}$ *when* r *and* l *are large and* $x + y = 1$

We introduce Gauss's Π-function, defined for real values of u greater than -1 by

$$\Pi(u) = \int_0^\infty e^{-t}\, t^u\, dt. \tag{1}$$

When u is an integer

$$\Pi(u) = u\,! \,. \tag{2}$$

Then

$${}^rC_l = \frac{\Pi(r)}{\Pi(l)\,\Pi(r-l)}. \tag{3}$$

When u is large, we have Stirling's approximation*

$$\Pi(u) = (2\pi)^{\frac{1}{2}}\, u^{u+\frac{1}{2}}\, e^{-u}\,\{1 + O(u^{-1})\}. \tag{4}$$

Let
$$l = rx + \alpha r^{\frac{1}{2}} + \eta, \tag{5}$$

where $|\eta|$ is less than a number k independent of r. Then

$$\Pi(l) = (2\pi)^{\frac{1}{2}}\,(rx + \alpha r^{\frac{1}{2}} + \eta)^{l+\frac{1}{2}}\, e^{-l}\,\{1 + O(r^{-1})\}, \tag{6}$$

$$\Pi(r-l) = (2\pi)^{\frac{1}{2}}\,(ry - \alpha r^{\frac{1}{2}} - \eta)^{r-l+\frac{1}{2}}\, e^{-r+l}\,\{1 + O(r^{-1})\}, \tag{7}$$

$$1/f(l) = \frac{\Pi(l)\,\Pi(r-l)}{\Pi(r)\,x^l y^{r-l}} = (2\pi r xy)^{\frac{1}{2}}\left(1 + \frac{\alpha}{xr^{\frac{1}{2}}} + \frac{\eta}{rx}\right)^{l+\frac{1}{2}} \times \left(1 - \frac{\alpha}{yr^{\frac{1}{2}}} - \frac{\eta}{ry}\right)^{r-l+\frac{1}{2}}\{1 + O(r^{-1})\}, \tag{8}$$

$$(l + \tfrac{1}{2}) \log\left(1 + \frac{\alpha}{xr^{\frac{1}{2}}} + \frac{\eta}{rx}\right) = (rx + \alpha r^{\frac{1}{2}} + \eta)\left\{\frac{\alpha}{xr^{\frac{1}{2}}} + \frac{\eta}{rx} - \frac{\alpha^2}{2x^2 r} + O(r^{-\frac{3}{2}})\right\} = \alpha r^{\frac{1}{2}} + \frac{\alpha^2}{2x} + \eta + O(r^{-\frac{1}{2}}), \tag{9}$$

* Cf. Whittaker and Watson, *Modern Analysis*, 12·33.

$$(r-l+\tfrac{1}{2})\log\left(1-\frac{\alpha}{yr^{\frac{1}{2}}}-\frac{\eta}{ry}\right)=-\alpha r^{\frac{1}{2}}+\frac{\alpha^2}{2y}-\eta+O(r^{-\frac{1}{2}}), \tag{10}$$

and therefore

$$\log 1/f(l)=\tfrac{1}{2}\log(2\pi rxy)+\frac{\alpha^2}{2xy}+O(r^{-\frac{1}{2}}), \tag{11}$$

so that

$$f(l)=(2\pi rxy)^{-\frac{1}{2}}\exp\left(-\frac{\alpha^2}{2xy}\right)\{1+O(r^{-\frac{1}{2}})\}^{*}. \tag{12}$$

II. *Approximation to* $g(l)={}^{r}C_l\,{}^{n-r}C_{m-l}/{}^{n}C_m$ *when* n, m, r *are large*

If we consider the expressions

$${}^{r}C_l\,x^l\,y^{r-l} \text{ and } {}^{n-r}C_{m-l}\,x^{m-l}\,y^{(n-r)-(m-l)},$$

where $x+y=1$ and x is arbitrary, we can choose x so that the maxima of both expressions occur for the same value of l, say l_0. It is necessary that

$$l_0=rx;\quad m-l_0=(n-r)\,x. \tag{1}$$

Hence
$$x=m/n;\quad y=(n-m)/n, \tag{2}$$

$$l_0=rm/n;\quad m-l_0=(n-r)\,m/n=(n-r)\,x,$$

$$r-l_0=r\,(n-m)/n=ry;$$

$$(n-r)-(m-l_0)=(n-r)\,(n-m)/n=(n-r)\,y. \tag{3}$$

Now put $l=l_0+p$. By Lemma I

$${}^{r}C_l\,x^l y^{r-l}=\{2\pi rxy\}^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\frac{p^2}{rxy}\right\}, \tag{4}$$

$${}^{n-r}C_{m-l}\,x^{m-l}\,y^{(n-r)-(m-l)}$$

$$=\{2\pi(n-r)\,xy\}^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\frac{p^2}{(n-r)\,xy}\right\}. \tag{5}$$

* I am indebted for this proof to Mr Newman; it replaces a somewhat longer one due to Bromwich, *Phil. Mag.* 38, 1919, 231–235.

Whence by multiplication

$$ {}^{r}C_{l}\,{}^{n-r}C_{m-l}\,x^{m}\,y^{n-m} $$

$$ = (2\pi xy)^{-1}\{r(n-r)\}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{np^2}{r(n-r)\,xy}\right\}. \quad (6) $$

But by (4) of Lemma I

$$ {}^{n}C_{m} = \frac{(2\pi)^{-\frac{1}{2}}\,n^{n+\frac{1}{2}}}{m^{m+\frac{1}{2}}\,(n-m)^{n-m+\frac{1}{2}}} = (2\pi nxy)^{-\frac{1}{2}}\,x^{-m}\,y^{-(n-m)}, \quad (7) $$

whence

$$ g(l) = \left(\frac{n}{2\pi xy\,r(n-r)}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{np^2}{r(n-r)\,xy}\right). \quad (8) $$

Consider now $$ \sum_{p=0}^{p} g(l). \quad (9) $$

When $r(n-r)\,xy/n$ is large this sum can be replaced approximately by an integral; when p increases by 1, one term is added to the sum. Hence

$$ \sum_{p=0}^{p} g(l) = \left(\frac{n}{2\pi xy\,r(n-r)}\right)^{\frac{1}{2}} \int_{0}^{p} \exp\left(-\frac{1}{2}\frac{np^2}{r(n-r)\,xy}\right) dp. \quad (10) $$

Put

$$ p = \{2r(n-r)\,xy/n\}^{\frac{1}{2}}\,\xi = \{2r(n-r)\,m(n-m)/n^{3}\}^{\frac{1}{2}}\,\xi. \quad (11) $$

Then

$$ \sum_{p=0}^{p} g(l) = \pi^{-\frac{1}{2}} \int_{0}^{\xi} e^{-\xi^2} d\xi \quad (12) $$

$$ = \tfrac{1}{2}\,\mathrm{erf}\,\xi, \quad (13) $$

where erf ξ denotes the error function defined by

$$ \mathrm{erf}\,\xi = \frac{2}{\sqrt{\pi}} \int_{0}^{\xi} e^{-t^2} dt. \quad (14) $$

Then (13) is the probability that l lies between rm/n and $rm/n + p$. The probability that it lies between 0 and r is unity; but since $1 - \mathrm{erf}\,\xi$ is insignificant for moderately

large positive values of ξ and $1 + \operatorname{erf}\xi$ insignificant for moderately large negative values of ξ, this is equivalent to

$$1 = \sum_{l=0}^{r} g(l) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\xi^2} d\xi,$$

which is true.

III. *Evaluation of*

$$\sum_{r=1}^{n} {}^{r}C_{l}\, {}^{n-r}C_{m-l} \quad \text{and} \quad \sum_{r=1}^{n} {}^{r}C_{l}\, {}^{n-r}C_{m-l}\, \frac{r-l}{n-m}$$

We have

$${}^{r}C_{l} = \frac{r!}{l!\,(r-l)!} = \frac{(l+1)(l+2)\ldots r}{1 \cdot 2 \cdot 3 \ldots (r-l)},$$

which is the coefficient of x^{r-l} in the binomial expansion of $(1-x)^{-(l+1)}$. Also, similarly, ${}^{n-r}C_{m-l}$ is the coefficient of $x^{(n-r)-(m-l)}$ in the binomial expansion of $(1-x)^{-(m-l+1)}$. Hence, by multiplying the two expansions, we see that $\sum_{r=1}^{n} {}^{r}C_{l}\, {}^{n-r}C_{m-l}$ is the coefficient of $x^{r-l} \cdot x^{(n-r)-(m-l)}$ in the expansion of $(1-x)^{-(m+2)}$; the coefficient, that is, of x^{n-m}. But this coefficient is

$$\frac{(m+2)(m+3)\ldots(n+1)}{1 \cdot 2 \ldots (n-m)} = {}^{n+1}C_{m+1} = \frac{(n+1)!}{(m+1)!\,(n-m)!}.$$

Also

$${}^{r}C_{l}\, \frac{r-l}{n-m} = \frac{r!}{l!\,(r-l)!}\, \frac{r-l}{n-m} = \frac{r!}{(l+1)!\,(r-l-1)!}\, \frac{l+1}{n-m}.$$

Whence

$$\begin{aligned} \sum_{r=1}^{n} {}^{r}C_{l}\, {}^{n-r}C_{m-l}\, \frac{r-l}{n-m} &= \frac{l+1}{n-m} \sum_{r=1}^{n} {}^{r}C_{l+1}\, {}^{n-r}C_{m-l} \\ &= \frac{l+1}{n-m}\, {}^{n+1}C_{m+2}, \text{ by the last result} \\ &= \frac{(l+1)(n+1)!}{(n-m)!\,(m+2)!}. \end{aligned}$$

INDEX

CAMBRIDGE: PRINTED BY W. LEWIS, M.A., AT THE UNIVERSITY PRESS